UCL DEPARTMENT OF GEOGRAPHY

POPULATION STRUCTURE AND THE SPATIAL ANALYSIS OF SURNAMES

James Allen Cheshire

Department of Geography, University College London

A thesis submitted in conformity with the requirements of Doctor of Philosophy (Ph.D)

September 2011

Word Count: 67, 175

DECLARATION

I, James Cheshire confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signed: Date:

ACKNOWLEDGEMENTS

I would most like to thank Professor Paul Longley, my primary supervisor, for his insights, guidance and high expectations from the moment I started this PhD. Taking on a fellow Southend boy from his rival school was always going to be a gamble – I hope that this thesis is evidence that it has paid off.

Thank you to Dr Pablo Mateos, my second supervisor, for his early ideas and helping me to settle in to my PhD at UCL. I am also grateful to Dr Alex Singleton for his practical advice. Additional thanks go to Daniel Lewis for his willingness to discuss ideas, Oliver O'Brien and Muhammad Adnan for their technical wizardry, and all three for ensuring that there has been at least one enjoyable coffee break per day.

I would like to acknowledge ESRI (UK) for providing financial support for my ESRC CASA Studentship. The success of this partnership is largely due to the efforts of Angela Baker, who has ensured that ESRI (UK) have helped in every way they can.

Extra special thanks go Isla Johns for her unrivalled support over the past three years (plus the four before that!). She has been with me every step of the way and reminded me that there is more to life than PhD research. My parents and brother have given me the confidence to pursue my chosen path in life and I hope I am a credit to them as a result. My final acknowledgement is to my late grandmother whose sense of humour and perceptiveness would have provided welcome breaks from PhD life and she is sorely missed.

THESIS OUTPUTS

PEER REVIEWED JOURNAL PUBLICATIONS[1]

2011 Identifying Spatial Concentrations of Surnames. ***International Journal of GIS*** (In Press). (J A Cheshire, P A Longley)

2011 People of the British Isles: Preliminary Analysis of Genotypes and Surnames in a UK Control Population. ***European Journal of Human Genetics*** (In Press). (B Winney *et al.*)

2011 Creating a Regional Geography of Britain through the Spatial Analysis of Surnames. ***Geoforum*** (In Press). Doi 10.1016/j.geoforum.2011.02.001 (P A Longley. J A Cheshire, P Mateos)

2010 The Surname Regions of Great Britain. ***Journal of Maps***. Doi 10.4113/jom.2010.1103. (J A Cheshire, P A Longley, A D Singleton)

2011 Delineating Europe's Cultural Regions: Population Structure and Surname Clustering. ***Human Biology.*** (Accepted) (J A Cheshire, P Mateos, P Longley)

PEER REVIEWED CONFERENCE PROCEEDINGS

2011 Spatial Concentrations of Surnames in Great Britain. ***Procedia Social and Behavioural Sciences*** (forthcoming). (J A Cheshire P A Longley).

2011 The Use of Consensus Clustering in Geodemographics. **Proceedings of *Geocomputation* 2011**, UCL. (J A Cheshire, M Adnan, C Gale)

2011 The Use of Consensus Clustering in Geodemographics. ***Proceedings of GIS Research UK 2011.*** GISRUK 2011, Portsmouth (J A Cheshire, M Adnan)

2010 Regionalisation and Clustering of Large Spatially-Referenced Population Datasets: the Case of Surnames. In Purves, R. and Weibel, R. (eds.)

[1] Papers published, or in press, are included in Appendix 4. Figures in this thesis included in the publications have been referenced accordingly.

Proceedings of GIScience 2010 available from http://www.giscience2010.org/**.** (J A Cheshire, P A Longley, P Mateos)

2010 Establishing Spatial Concentrations of Surnames Using Kernel Density Estimation. In De Felice, S. *et al.* (eds.) ***Proceedings of the GIScience 2010 Doctoral Colloquium***. IFGI Print Series, Heidelberg, Germany. (J A Cheshire)

2010 Informing Population Genetics through Spatial Analysis of Surnames. In Haklay, M., Morely, J. and Rahamtulla, H. (eds.) ***Proceedings of GIS Research UK 2010. GISRUK 2010***. UCL, London. (J A Cheshire, P A Longley, P Mateos)

2010 Surnames as Indicators of Cultural and Linguistic Regions in Europe. In Haklay, M., Morely, J. and Rahamtulla, H. (eds.) ***Proceedings of GIS Research UK 2010. GISRUK 2010***. UCL, London. (J A Cheshire, P A Longley, P Mateos)

2009 Surnames as Indicators of Cultural Regions in 2009 the UK. In Fairbairn, D.(ed.) ***Proceedings of GIS Research UK 2009. GISRUK 2009.*** University of Durham, Durham**.** (J A Cheshire, P Mateos, P A Longley)

2009 Combining Historic Interpretations of the Great Britain Population with Contemporary Spatial Analysis: The Case of Surnames**.** In ***Proceedings of the IEEE Geospatial Computing Workshop December 2009.*** (J A Cheshire)

OTHER CONFERENCE PRESENTATIONS

2011 The Use of Surnames and Spatial Analysis: Utilising a Novel Dataset to Map Population Change. ***Association of American Geographers Annual Meeting***. Seattle, USA. (J A Cheshire)

2010 The Use of Multidimensional Scaling to Inform the Colour Values of a Choropleth Map. ***British Cartographic Society Annual Conference.*** Nottingham, UK. (J A Cheshire)

2010 Spatial Structure in Surname Distributions: Establishing Cultural Regions in Great Britain. ***Cultural Evolution of Spatially Structured Populations.*** UCL, London, UK. (J A Cheshire, P A Longley)

2010 Visualising Social Change: The Case of Surnames. ***Royal Geographical Society Annual Conference.*** London, UK. (J A Cheshire)

2009 Surnames as indicators of Cultural Regions in Great Britain. *Regional Science.* Limerick. (J A Cheshire)

2009 Surnames: A Rich Source of Geodemographic Data. ***Royal Geographical Society Annual Conference.*** University of Manchester, Manchester, UK. (J A Cheshire, P A Longley, P Mateos)

2009 Surnames as indicators of Cultural Regions in Great Britain. ***Popfest***. London School of Economics, London, UK. (J A Cheshire, P A Longley, P Mateos)

2009 Surnames as Indicators of Cultural Regions: Mapping the "Population Geology" of the UK. ***S4 European Spatial Analysis Network Conference.*** UCL, London, UK. (J A Cheshire)

INVITED PRESENTATIONS

2011 Spatial Analysis and Visualisation Using Free Data. ***ESRI (UK) Technical Presentation***. ESRI (UK) HQ, Aylesbury, UK.

2010 Global Surname Mapping. **University of Nanjing**, Nanjing, China.

2010 Spatial Structure in Surname Distributions. ***CASA Seminar.*** UCL, London, UK.

2010 The Geo-Genealogy of British Surnames. ***British Society of Population Studies Day Meeting***. University of Cambridge, Cambridge, UK.

2010 Surnames and Spatial Analysis. ***University of Leicester Department of Genetics***. Leicester, UK.

2009 Developments in GIS at UCL. ***ESRI (UK) Technical Presentation.*** ESRI (UK) HQ, Aylesbury, UK.

2009 Surnames as Indicators of Cultural Regions in the UK. ***CASA Seminar***. UCL, London, UK.

BOOK REVIEWS

2010 Fischer, M. and Getis, A. (eds) Handbook of Applied Spatial Analysis. ***Environment and Planning B: Planning and Design***. 37, 6: 1140-1141.

2010 Rogerson, P. and Yamada, I. Statistical Detection and Surveillance of Geographic Clusters. ***Environment and Planning B: Planning and Design***. 37, 4: 761-762.

2010 Hoff, P. A First Course in Bayesian Statistical Methods. ***Journal of the Royal Statistical Society: Series A***. 173, 3. 694-695.

2009 Sarkar, D. Lattice: Multivariate Visualisation with R Review**.** ***Journal of the Royal Statistical Society: Series A***. 173, 1. 275-276.

2009 Bivand, R., Pebesma, E. and Gómez-Rubio, V. Applied Spatial Data Analysis with R. ***Significance.*** 6, 3: 138-139.

PRIZES AND AWARDS

2010 ***Early Career Travel Bursary*** for an exemplary paper, GIScience Conference (5 prizes, 25 entries).

2010 ***Student of the Year***: UCL ESRI Development Center.

2010 ***2nd Place***: UCL Graduate School Poster Competition (15 Entries).

2009 ***Best Young Researcher***: Regional Science Annual Conference (British and Irish Section).

ABSTRACT

Geographers have largely overlooked surnames (family names), and their geographic concentrations, as a valuable source of data to indicate the spatial structure of populations. This thesis seeks to provide a substantive contribution to the geographical literature by demonstrating how quantitative spatial analysis of surname data can be used as an aid to understanding population structure at a range of scales from the regional to the continental. The primary purpose of this research is not to develop detailed case studies or to investigate specific examples of population characteristics considered interesting for their novelty: rather, the core concern is to focus on the identification or confirmation of generalised trends. Much of the current research that uses surnames (for example in population genetics) contains a geographical element, yet stops short of exploiting and accommodating the effect of scale, shape and size of spatial units. The application of computationally intensive spatial analysis techniques to a comprehensive and innovative dataset (see worldnames.publicprofiler.org) makes it possible to address these issues for the first time. The thesis develops and applies a robust analytical and methodological framework for the analysis of surnames as a primary data source. Applications of the research are used to demonstrate the utility of surnames in studies of population genetics, in migration research, as well as in the spatial analysis of large datasets more generally.

TABLE OF CONTENTS

LIST OF FIGURES[2]

[2] *All boundary data used for maps of Great Britain is Crown Copyright Ordnance Survey 2011.*

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	Definition
AUC	Area Under Curve
CAS	Census Area Statistics
CDF	Cumulative Density Function
GISc	Geographic Information Science
GOR	Government Office Region
IDW	Inverse Distance Weighting
KDE	Kernel Density Estimation
LSOA	Lower Super Output Area
LQ	Location Quotient
MAUP	Modifiable Areal Units Problem
MBA	Monmonier's Barrier Algorithm
MDS	Multidimensional Scaling
MLQ	Maximum Location Quotient
MSOA	Middle Super Output Area
NSPD	National Statistics Postcode Directory
NUTS	Nomenclature of Units for Territorial Statistics
OA	Output Area
OAC	Output Area Classification
ONS	Office for National Statistics
OS	Ordnance Survey (Great Britain)
PAM	Partitioning Around Medoids
R	R Project for Statistical Computing and Graphics
RGB	Red, Green, Blue colour model
SOA	Super Output Area
SOMs	Self-Organising Maps
SRTM	Shuttle Radar Topography Mission

UK	United Kingdom of Great Britain and Northern Ireland
Withinss	Within Sum of Squares

"It may be thought by some that the investigation of the distribution of names is an idle amusement, productive of no utility of man. I have come to think, however...that it is a matter of much importance to the antiquarian, the historian the ethnologist and also to the more practical politician"

Henry Guppy 1890: vi.

1 INTRODUCTION

Family names, or surnames, provide almost every culture with a ubiquitous method for distinguishing between familial groups. On a daily basis surnames are required as a means of identification by a wide range of organisations and individuals. Yet the routine use of surnames has meant that their cultural and geographical significance is often overlooked. We often make judgements about a person's likely ancestry based on their surname and often (either consciously or subconsciously) assign them to a country or linguistic group as a result. It is obvious, for example, that surnames such as "Smith", "Jones" and "Macleod" are English, Welsh and Scottish in origin, respectively. Placing surnames within a regional context becomes somewhat more specialist yet many people would, for example, relate the surname "Crosby" to the town of the same name in Lancashire, England. Even within cities it would not be unreasonable to suggest that a person named "Cohen" is likely to live in one of the Jewish areas of London such as Golders Green or Stanford Hill. From such anecdotes alone, it is clear that many surnames contain spatial information at a variety of scales relating to the origins, and probable areas of residence, for many of their bearers. The purpose here is to move beyond such anecdotal conjecture to provide unprecedented insights into the geography of surnames.

This thesis seeks to provide a substantive contribution to previous studies of surnames, and the geographical literature more widely, by demonstrating how spatial analysis and surname data can be used to effectively decipher population structure. It does so at a range of scales from the sub-national to continental and for two time periods. The analysis is applicable to both people, in the sense that individual surnames shed light on ancestry and relatedness, and places, based on the insights provided by the surname composition of particular areas. The datasets used are unprecedented in terms of their detail and extent and have required a large amount of computational storage and processing to generate the results outlined in Chapters 3 to 6. Beyond advances in data and analysis, this thesis also seeks to begin to rectify the lack of interest in surnames from the geographical literature by placing previous research, undertaken primarily by population geneticists, in the context of the past

five decades of quantitative geography research. For the first time, the impacts of well-known issues concerning spatial data are considered for surname analysis. It is hoped that both the comprehensiveness of the analysis and the thought given to more conceptual issues will provide a firm foundation to future surnames research in geography and population studies more widely.

As the aims, outlined below, demonstrate, this thesis is not seeking to create a wide-ranging gallery of applications to showcase surnames as a useful source of quantitative spatial data. Such an approach, the results of previous research suggests, would not amount to a systematic treatment of both data or methods. The purpose of this thesis, therefore, is to develop the foundations required by advancing previous research (primarily within population genetics) to offer a set of robust methodological and analytical insights capable of adequately capturing the spatial characteristics of surnames. A number of applications are suggested related to studies of migration and population genetics, although it is recognised that they merely scratch the surface of the multifaceted contributions that analysis of surnames can make to a multitude of fields.

1.1 AIMS

As the title of this thesis suggests, the general purpose here is to investigate the ways in which surnames can be used to unearth, primarily spatial, structure within populations. Surnames are spatially non-random phenomena, with the majority continuing to be concentrated in or near the areas where they first appeared many generations ago. In the more recent past, name concentrations can also indicate the first destinations of international migrants to a country. If individuals frequently undertook long-distance migrations to random locations then population structure in this sense would have eroded to the extent that it is no longer measurable. Migration nonetheless remains an experience that the majority of people seek to avoid. For this reason many population traits, such as dialects or surnames, vary systematically across space. These variations can be unearthed through a number of quantitative measures that characterise the spatial distribution of surnames both in terms of single

surnames and also broader regional patterns. On this basis the thesis has four broad aims:

1. To review previous surnames research in the context of spatial analysis and quantitative geography more broadly.
2. To create a methodology for the automated identification of key characteristics pertaining to individual surnames, such as extent and area(s) of highest concentration.
3. To review and establish the methodological processes required for the aggregation and regionalisation of surnames appropriate to a range of scales and data sources.
4. To promote the outcomes from 2 and 3 as a basis for future research and hypothesis generation relevant to both a range of applications and disciplines.

The purpose of this thesis is not to highlight anomalies or investigate specific examples of population characteristics considered interesting for their novelty. Instead, the above aims will be fulfilled through a focus upon what Pooley and Turnbull (1998: 330) describe as the "everyday and commonplace dimensions of population movement" implicit in surnames.

The substantive focus of this thesis (aims 2 and 3) concern the meaningful aggregation and generalisation of surname distributions from the rich surname datasets compiled by UCL Department of Geography (see worldnames.publicprofiler.org and gbnames.publicprofiler.org). The surnames of Great Britain are given the most detailed treatment in this thesis, reflecting the interests and expertise of the author in addition to the availability of both historical and contemporary data. The methods successfully developed in the British context are extended and applied to 15 European countries in Chapter 6, in order to demonstrate the wider geographical relevance of the research. A more detailed outline of the thesis is provided below.

1.2 THESIS STRUCTURE

1.2.1 CHAPTER 2: SURNAMES AS SPATIAL DATA

Chapter 2 provides the research background and context for the thesis. It outlines the historical, linguistic and cultural significance of surnames, enforcing their validity as phenomena worthy of research. This is followed by an overview of previous surnames research with special reference to its shortcomings in the treatment of surnames as spatial data. The purpose of this chapter is to provide the foundations upon which subsequent chapters in this thesis are built.

1.2.2 CHAPTER 3: SURNAME DATA AND PRELIMINARY ANALYSIS

Chapter 3 describes the datasets used here, namely the 1881 Census of Great Britain, the 2001 Enhanced Electoral Register for Great Britain and the surnames and locations for almost 150 million individuals from 16 European countries taken from the UCL Geography Worldnames database. An appraisal of the quality of the datasets used and their representativeness of the target populations is provided alongside spatial referencing information. In addition, the chapter includes some preliminary analysis of the data for Great Britain. Its purpose is to offer insights into the spatial distributions of surnames at both the individual and aggregate level. Such insights provide the context to many of the processes captured in the more substantive analysis undertaken in Chapters 4 to 6. It also serves to enforce and advance many of the results outlined in Chapter 2 from other studies investigating the spatial distributions of surnames.

1.2.3 CHAPTER 4: TOOLS TO DISCERN SPATIAL PATTERN: DETECTING SURNAME CLUSTERS

Chapter 4 is the first of the substantive chapters. It addresses the specific need for a systematic, comprehensive and automated method for discerning the spatial

characteristics of individual surnames. The approach taken treats the discovery of areas of highest concentration in a surname's distribution as a spatial clustering problem. It provides a review of a number of potential methods appropriate to the task before settling on kernel density estimation (KDE) for the final methodology. Details of the methodological steps undertaken to create a surname typology are explained in detail before a selection of results are discussed. The final aspect of Chapter 4 explores the temporal aspects of the analysis and how it can be used to chart historical population processes.

1.2.4 CHAPTER 5: AGGREGATION AND REGIONALISATION

Chapters 5 and 6 both combine methods and interpretations from the fields of population genetics and quantitative geography through a focus on adequate measures of surname relatedness - or surname distance - between localities or regions and areal classification algorithms to partition space according to such distances. The resulting regions are a statement of within-region similarities and between region differences. Chapter 5 begins by reviewing why regions are a valid conceptualisation of population data, before introducing the methods used for the creation of surname regions for Great Britain. The effectiveness of these methods is then discussed to lay the foundations of Chapter 6.

1.2.5 CHAPTER 6: APPLICATIONS AND EXTENSIONS OF SURNAME REGIONS

Chapter 6 is focussed on applying and extending the methodological insights provided by Chapter 5. The first section demonstrates the utility of surname regions in unearthing past migration and outlines the ways in which contemporary analysis can be compared to both historical boundaries and historical research. The second, more substantive, aspect of Chapter 6 seeks to provide both a methodological and geographic extension to the analysis of Great Britain's surname regions outlined in Chapter 5. It does so by detailing the first investigation of geographical structure of surnames at a continental level covering 16 European countries, in addition to

proposing a clustering technique appropriate for the pan-European scale. The results are a regionalisation of Europe based purely on the geography of surname frequencies. Chapter 6 concludes with a general discussion concerning the merits of creating a regional geography based on surnames.

1.2.6 CHAPTER 7: METHODOLOGICAL CONTRIBUTIONS, APPLICATIONS AND RESEARCH PROSPECTS

It is in the context of the innovative, robust and analytical practices outlined in the previous chapters that Chapter 7 seeks to consolidate the achievements of the thesis and identify a path for future developments. It outlines the methodological contributions to surnames research and spatial analysis more generally before demonstrating the relevance of the research undertaken for population studies, including population genetics. The chapter ends by exploring potential avenues for future research.

1.2.7 CHAPTER 8: THESIS SUMMARY AND CONCLUSIONS

Chapter 8 draws together the multiple research strands of this thesis to assess the extent to which they address the four aims stated in the introduction. Each aim is re-stated and discussed before final conclusions are offered.

1.2.8 NOTE ON SOFTWARE AND ASSOCIATED CODE

The majority of the analysis has been undertaken in the R Software Environment for Statistical Computing and Graphics (R Development Core Team 2011). This is an open source program freely downloadable from www.r-project.org and the code used to produce many of the outputs in thesis is available on request. The two other software packages used here are ArcGIS 10 and MySQL and any code used for these can also be requested from the author.

2 SURNAMES AS SPATIAL DATA

Surname adoption did not occur simultaneously in all inhabited places and surnaming conventions have always been a product of both cultural (including linguistic) and legislative processes. Such processes are systematic but not geographically uniform, resulting in spatial structuring of surname distributions that may nevertheless subsequently be obscured by population movements. This chapter is concerned with the conceptualisation of surnames as spatial and quantitative indicators of cultural, linguistic and genetic characteristics. The purpose here is not to provide an in depth study of the history and continuing development of surnames in Great Britain and Europe, although examples will be used to demonstrate the utility and possibility of using surnames as indicators of cultural, linguistic and genetic diversity. Rather, the purpose of this chapter is to provide a rationale for the spatial analysis of surnames.

The first section of the chapter will outline the historical, linguistic and cultural significance of surnames and the gravitas that these associations lend to surnames as a tool for researching populations. The second section addresses previous research into the spatial distributions of surnames. Much of this provides the foundations on which subsequent chapters in this thesis are built. The field of population genetics is arguably the largest contributor to the study of surnames and is the focus of the third section. The link between surnames and genetics adds further weight to the utility of surname analysis and alludes to a number of other cultural processes that are not fully considered in the context of surnames research. The penultimate section is concerned with the special features of spatial data that have been overlooked in previous research; this leads into the final aspect of the chapter, which outlines a number of research needs that relate to the motivations for the analysis undertaken in this thesis.

2.1 SURNAMES, CULTURE AND LANGUAGE

2.1.1 SURNAME ORIGINS

Surnames, culture and language are closely linked so it follows that people are unlikely to assign unpronounceable or unfamiliar names to themselves or the people they interact with. In Britain, surnames were developed as a means to distinguish individuals (particularly men) from one another at a time when there were very few forenames. For example, the poll-tax return of 1379 from the Midlands contains records of 715 men with only 22 forenames between them; over 50% were called John or William (Hey 2000).

The British case is typical of many countries in Europe that experienced gradual surname adoption processes sometime in the last millennium. Whilst it is unclear precisely when surnames became formalised and hereditary in Great Britain (Barker *et al.* 2007), there is agreement that the Doomsday Book of 1085 made surnames a necessary (but not compulsory or hereditary) method of distinguishing between individuals (Barker *et al.* 2007). The lack of any legal basis to surname adoption has led to the view that surnames were acquired gradually across the population. In the 13th Century, surnames closely allied to locality were being regularly recorded; however, these were unlikely to be hereditary (McClure 1979). By the 15th Century inheritable surnames became generally adopted in England (Lasker and Mascie-Taylor 1985), but it was not until the 16th Century that they were fully adopted in Scotland (Barker *et al.* 2007).

Fortunately, implicit in the surnames themselves is a much clearer indication of the conventions and inspirations people used when selecting them. The majority of surnames can be categorised into local surnames, occupational surnames, surnames of relationship, or nicknames (Barker *et al.* 2007). Preference for each naming convention varies geographically; in Wales, for example, the trend was towards patronyms (surnames taken from a father's forename), while the people of Sussex opted for toponyms (surnames derived from place names) (Hey 2000). Toponyms are classified as locative (taken from specific places) and topographical (from general landscape features) and are the most obvious link between surnames and geography. Table 2-1 provides some examples of common naming conventions in Great Britain.

Table 2-1: A categorisation of British surnames. Adapted from Barker *et al.* (2007: 15).

Category	Example	Explanation
Occupational (Metonyms)		
Profession	Smith	Blacksmith/ metal worker
Office/ Trade	Reeve	Chief magistrate/ overseer
Rank/Status	Knight	A knighted person
Occupation Features	Falconer	One who kept/trained Falcons
Local Surnames (50% of		
Toponymic (Topographic)	Rivers	Dweller near river
Toponymic Locative)	Cornwall	Man from Cornwall
Habitation (residence)	Gate	Habitation at/near a gate
Habitation (work)	Hall	A worker at the hall.
Surnames of Relationship		
From personal name	Jones	Son of John
From personal name	Margaretson	Son of Margaret
Personal name from other relative	Johnson	Related to John
Personal name from diminutive	Dickens	Son of Dick (Richard)
Clan or tribal names	MacBain	Related to the MacBain clan.
Nicknames		
From animals	Fox	Slyness or other attributes
From characteristic traits	Careless	Free from care/ responsibility
From objects	Shorthose	Someone who wore short boots
From physical features	Little	A small person
From times and seasons	Pasque	Person born at Easter
From iconic description	Drinkwater	Heavy drinker

2.1.2 SURNAMES AND INHERITANCE

The inheritable nature of surnames is extremely important because it creates a self-maintaining, enduring and culturally significant surname geography. Patterns of inheritance form the basis for genealogical research and enable the creation of familial lineages that are often traceable to the point in time when the surname was first formalised. It has not been possible to explain fully why surnames became paternally inherited (Hey 2000). One plausible explanation is simply that surnames originated as an indicator of land ownership which, like many assets, was passed from father to son. The earliest adopters of surnames in Poland (around the 15th Century), for example, were the nobility, as a statement of the villages they "owned". As these responsibilities were generally passed from fathers to sons, the surnames went with them. This was also the case in Great Britain, where it was common for Norman barons to continue to possess surnames associated with the estates they left in Northern France; "Mortimer" for instance is derived from the town of Mortemer in Normandy (Hey 2000).

A surname's path down the male line is broken if there are no sons to pass it on to or if the name is changed during marriage or for other reasons, such as changing fashions. In the latter case, for example, there are no longer people called "Smelly" in Great Britain (according to the 2001 Electoral Roll) but they were present (and probably in decline!) in 1881. Based on the small populations associated with surnames that existed in 1881 but not 2001, the decline of surnames in this way is not considered problematic for the analysis. A break in lineage due to a lack of male heirs is a more significant phenomenon and one that is hard to quantify in the context of this research. Change through marriage (from a male perspective) continues to be unlikely.

2.1.3 WOMEN AND SURNAMES

One of the largest (and perhaps most obvious) limitations of this research is that surnames reflect a male-oriented view of the world. The fact that the majority of

women change their surname when they marry means that they adopt the same cultural marker as their partner. This creates minor problems in the classification of areas but more significant problems in the classification of people.

As Table 2-2 shows, in many European countries women adopt their partner's surname after marriage. The surname may represent an entirely different culture, and of course, cannot be used to apply the assumption that women who share the same surname are more likely to be related than those who do not. When classifying an *area* (as opposed to an individual) based on its surname composition, the inclusion of female surnames is less problematic because they could either have adopted a surname from a local male, or still bear a surname inherited from their father. The former case will be surprisingly common in the dataset given that a study of marriages in 1979 found that over half were between couples living (and most likely born) within 5km of each other (Coleman and Haskey 1986). The distribution of "marital distance" is shown in Figure 2-1. Such conclusions are also true at a European level, and certainly in rural areas, with many matrimonial migrations being less than a few kilometres (Manni *et al.* 2008).

Table 2-2: The proportion of men and women who change their surname after marriage for a selection of European countries.

	Practices				Opinions						
	Which surname do you use? Which surname does your wife use?				It is preferable for the wife to use:					**Men** Opinions favourable	**Women**
	The husband's surname	**Both**	**The wife's surname**	**Other**	**Her husband's surname**	**Both surnames**	**Her own surname**	**Other**	**No opinion**	to using the surname:	
										of the husband	
Germany	**95**	3	1	< 1	**64**	16	1	11	8	68	62
United Kingdom	**94**	4	1	< 1	**71**	15	1	9	3	70	72
Austria	**93**	4	2	2	**67**	17	1	10	5	70	64
France	**91**	7	2	0	**49**	40	2	6	3	49	49
Ireland	**90**	7	2	2	**59**	22	4	7	8	60	58
Sweden	**87**	7	6	< 1	**53**	21	4	17	5	54	52
Denmark	**71**	13	13	2	**37**	29	10	18	6	35	38
Netherlands	**55**	41	4	< 1	**41**	39	7	9	4	39	42
										of both husband and wife	
Luxembourg	41	**47**	4	8	31	**49**	5	11	5	46	52
Belgium	22	**57**	20	1	21	**60**	13	1	5	58	62
Italy	12	**64**	21	2	13	**68**	16	1	2	66	69
										of the wife	
Spain	4	17	**77**	2	8	23	**59**	5	5	51	66

Field: married men and women aged 18 and over • Source: Eurobarometer survey, 1995•Adapted from Valetas 2001.

It would be impossible to account for all the exceptions to females adopting their spouse's surnames (and in some cases vice versa) and so for this thesis it was considered best to treat surnames with male and female bearers in the same way. Aside from the uncertainty inherent in assigning genders to individuals based on their forenames, a strength of this research is the use of the most complete population data available and adaptation of it would result in unnecessary information loss. That said, some degree of sampling is inherent in the European datasets outlined in Section 3.1.3 obtained from telephone directories, as entries are widely acknowledged to be predominantly male household heads.

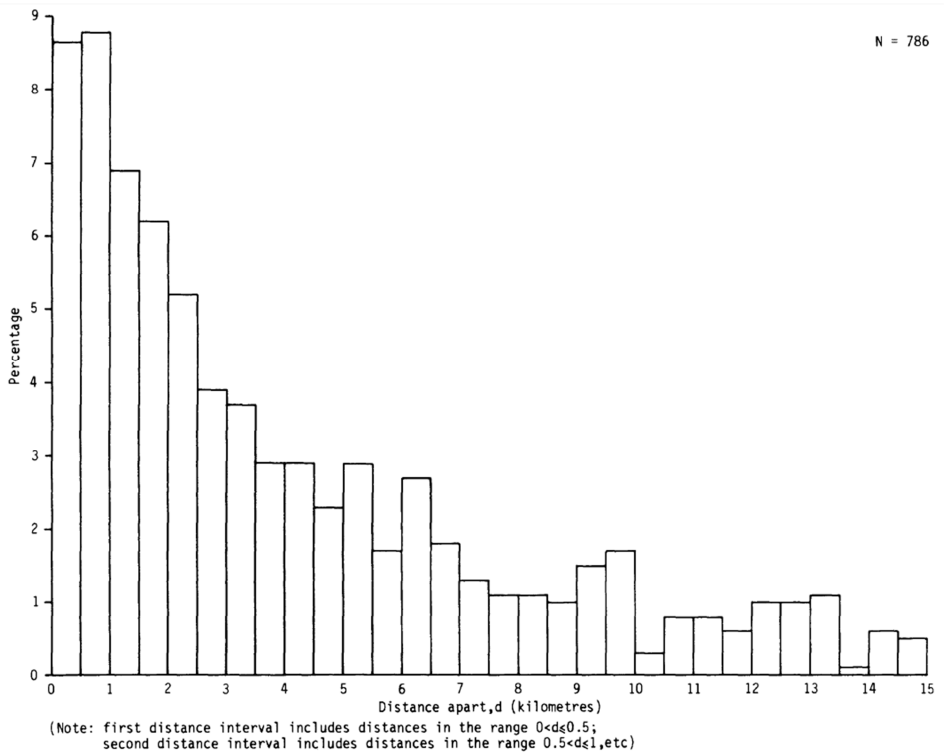


Figure 2-1: The Distribution of "Marital Distances" (in kilometres) between the individuals of 786 couples. Taken from Coleman and Haskey (1979: 344).

2.2 SURNAMES AND GEOGRAPHY

As is demonstrated below, and in the preliminary analysis outlined in Chapter 3, surnames are not geographically random phenomena and therefore provide an interesting source of data for spatial analysis. In spite of this, studies of surname frequency distributions have been rare in the geography literature (Zelinsky 1997) and the few published examples, such as Porteous (1982), relate to one or a small number of specific surnames and cannot be representative of the broader population.

One of the earliest attempts to define surname regions was undertaken by Guppy (1890) in his book "Homes of Family Names in Great Britain". The book, which remains one of the most comprehensive, sought to conceptualise surnames as regional phenomena. Many of the issues, such as whether the Welsh border defines the extent of Welsh communities, or whether Parliamentary areas "are not political or artificial" in their determination (Guppy 1890), are still important today. Table 2-3 contains the surname categorisation developed by Guppy. From his classification Guppy established that clear regions existed: South West England's inhabitants, for instance, possessed 40% of all 'peculiar' names. Inspired by the historical regions of Anglo Saxon Britain, Guppy suggested that the regionality of surnames could be sufficient to restore "the heptarchy to our land". As will be demonstrated in Chapter 6, Guppy's work was extremely perceptive, with his entirely manual approach reaching similar conclusions to the contemporary computational approaches used here.

Table 2-3: Guppy's classification of British surnames. These categories are still applicable to many contemporary surnames. Taken from Guppy (1880: 11).

Classification	Occurrence
General Names	30- 40 Counties
Common Names	20- 29 Counties
Regional Names	10 - 19 Counties
District Names	4- 9 Counties
County Names	2 – 3 Counties (principle home in one of them)
Peculiar Names	1 County (and generally to a specific parish/ division within it.)

More recently, Zelinsky (1970) used forenames as a data source to investigate cultural variation across 16 counties in the Eastern United States but no further work appears in the geographical literature until Porteous (1982). This study suggests a multi-operational method for investigating the spatial origins and subsequent diffusion of rarer English surnames at a national and regional scale (Porteous 1982). Despite Porteous' assertion that "names have been neglected by geographers" (Porteous 1982: 395), and his attempt to reintroduce surname studies to geography, the article failed to stir much interest. Zelinsky (1997) also unsuccessfully encouraged geographers to use people's names in the study of population and regions. He argued that the lack of geographical interest in the study of surnames remained a mystery with "only occasional probes" (p. 465) into the links between names and wider questions in the humanities and social sciences. The latest studies in the geographical literature attempt to do this and, like this thesis, result from research undertaken at UCL Geography (See Longley *et al.* 2007, Longley *et al.* 2011a).

Fields such as human biology, population genetics and linguistics are the largest contributors to the "thousand-fold" increase in the number of publications (see Colantonio *et al.* (2003: 785)). As will become apparent, the majority of these studies have a spatial element but are limited by a lack of comprehensive data or an appreciation of some well-studied issues in spatial analysis. What follows is a review of published research related to the geography and spatial distribution of surnames at a variety of scales.

2.2.1 PREVIOUS RESEARCH

The majority of research into the spatial distributions of surnames has been undertaken by population geneticists, whose studies have used small samples of names and have been highly focused on issues pertaining to genetics rather than spatial analysis. A reliance on sampling is representative of the limited availability of comprehensive, spatially referenced population datasets in digital form, besides the necessary computer power or manual resources to process them. This review will

focus first on the smaller scale studies of Anglo-Saxon surnames before discussing some more generalized research at European level.

Research into Anglo-Saxon surnames has been either descriptive, by simply outlining surname origins or meanings, or concerned with levels of isonymy (sharing the same name) between or within populations. The latter is more important in the context of this research and was pioneered by Lasker during the 1970s and 1980s (Lasker 1985). Lasker and colleagues selected surnames from telephone directories or marriage records to undertake studies of both regional (for example in Henley-on-Thames (Fox and Lasker 1983) or Oxfordshire (Lasker 1999)) and national level (see Mascie-Taylor and Lasker 1985, Lasker 1985). This work represents among the first and most comprehensive attempts to use surnames as a quantitative data source in Britain.

In England, Kaplan and Lasker (1983) found almost twice the expected number of surnames located in areas that provided location specific toponymic names. Although some of the surnames (taken from 1981 English telephone directories) only partially originated from the studied areas, a tendency of association appeared to remain, despite the long period since surname establishment (Kaplan and Lasker 1983). This study also found a clear distinction between large urban areas and the rest of England. For example, they observed 10 times the expected population of "Pocklingtons" in the town of Pocklington (Yorkshire), and 4 times the number of "Dudleys" in Dudley (West Midlands) but fewer than half the expected frequency of "Birminghams" in Birmingham and a sixth the number of "Sheffields" in "Sheffield" (Kaplan and Lasker 1983). The work also demonstrates the attenuating effect of distance on surname concentrations as they reduce with distance from the source area. An illustration of this point is the finding that surnames derived from place-names within a 50km radius of Manchester and Birmingham occur at 145% of the expected frequency, reducing to 124% 50-99km away and 82% of expected over 150km away (Kaplan and Lasker 1983).

Beyond research into specific regions and samples of surnames, few studies have sought to create a national picture of the surname distributions in Great Britain. As

has already been discussed above, conjecture surrounding the existence and extent of Britain's surname regions dates back as far as Guppy (1890). He sampled around 8,000 surnames to partition Great Britain into seven surname regions. These regions were subjectively established from Guppy's experience in manually compiling (through placing buttons on a map) the spatial distributions of each of the surnames in his book.

There was no further research into the regionalisation of surnames until the more substantive applications of quantitative analysis developed by geneticists (see Section 5.2) and facilitated by digital databases and increased computational power. The maps in Mascie-Taylor *et al.* (1985) marked a revival for the national-level quantitative analysis of surnames. They depict the geography of some of the most common British surnames by mapping surname frequency alongside plots of the probability of local excess or deficiency from north to south and east to west. In addition, these maps were used as an approximate comparison to the descriptions provided by Guppy (1890). This, like many other studies, such as Smith *et al.* (1984), provides very superficial comparisons and perhaps reflects a lack of digital data. More comprehensive and geographically extensive analysis of surnames geography is now available online for Great Britain (gbnames.publicprofiler.org) and 26 countries worldwide (worldnames.publicprofiler.org). These websites, created by UCL Department of Geography, enable anyone with Internet access to visualise the geographies of individual names as well as selected ethnic groups associated with the origin of those same names. By utilising the same data, this thesis is the first to provide comprehensive historical comparisons of British surname distributions using the "complete" population record provided by the 1881 Census.

Whilst interesting from a genealogical perspective, choropleth maps of individual surnames, such as those produced by the above websites, provide only a limited picture of the milieu in which they were first coined. Sokal *et al.* (1992) sample 100 surnames in England and Wales (the 84 most common names combined with "some selected rarer names" (p. 447)) and use surface wombling (see Barbujani and Sokal. 1990) to derive surname frequency boundaries. They, like Kaplan and Lasker (1983), produce strong evidence of isolation by distance; that is, populations further apart

are less likely to mix and share surnames. Their study nonetheless is limited because it used only a sample of surnames using frequencies taken from recorded marriages that occurred between January and March 1975. These can only be representative of a small proportion of the English and Welsh population (Scotland was not included in the study). Despite this limitation, Sokal *et al.* (1992) posit 21 abrupt changes in surname compositions between English and Welsh marriage registration districts. It is clear that Anglo-Saxon surnames and their distribution in Great Britain have provided a productive area of research, the conclusions from which are largely applicable to other countries.

Comparable European studies have been ongoing for a similar period. Since the early work of Cavalli-Sforza and colleagues using Italian telephone directories on magnetic tape in the 1970s (Cavalli-Sforza *et al.* 2004), and specifically their wide availability in CD format in the 1990s, a host of studies have begun to analyse the surname structure of populations at national level. One group has dominated this research through the publication of a variety of national-level papers for surnames. European examples include: Austria (Barrai *et al.* 2000); Switzerland (Rodriguez-Larralde *et al.* 1998); Germany (Rodriguez-Larralde *et al.* 1998); Italy (Manni and Barrai 2001); Spain (Rodriguez-Larralde *et al.* 2003); Belgium (Barrai *et al.* 2004); the Netherlands (Manni *et al.* 2006); and France (Scapoli *et al.* 2005). All studies demonstrate the spatial structure of surnames in the countries studied, with fewer commonalities between populations the further away they were. Such differences are thought to be largely indicative of the different linguistic and cultural histories within or between the countries studied.

As mentioned above, people were unlikely to assign themselves unpronounceable surnames. This has been confirmed, on the European level at least, by a number of studies. In France (Scapoli *et al.* 2005) and Belgium (Barrai *et al.* 2003) the dialect transitions closely match those of surnames, meaning that measures of each can be used interchangeably. In the Netherlands, however, Manni *et al.* (2008) found no statistically significant relationship between surnames and language. This finding is interesting given the dialects of the country and suggests that other factors can influence the surnames chosen. In the case of the Netherlands, Manni *et al.* (2008) cite religion as a possible explanation with surname transitions occurring along the

border between Protestant and Roman Catholic areas. The extent to which such differences (between surnames and language) are visible depends on the scale and linguistic diversity of the populations studied. If establishing broad surname regions in Europe, as Scapoli *et al.* (2007) have done, it is clear that language is the key determinant.

The Scapoli *et al.* (2007) study uses data from 8 European countries to create a continental-level surname regionalisation. It, like those listed above, was reliant on sampling and only included data for 2094 towns and cities grouped into 125 spatial units. Scapoli *et al.* (2007) found clear regionalisation patterns in surname compositions, closely matching the national borders for the eight countries included with exceptions showing the geo-historical distribution of languages. Whilst being exceptionally extensive, both geographically and in terms of the number of surnames sampled, the work by Scapoli *et al.* (2007) is still limited by its partial sampling of "representative" locations. This thesis marks a significant extension to this work through the inclusion of 8 more European countries (16 in total), and incorporating "complete" populations and locations (that is, without sampling).

2.3 SURNAMES AND GENETICS

The use of 'cultural traits' (as they are known in human biology) such as surnames, to suggest genetic differences between population groups has become widely accepted. Cultural traits are transmitted from ancestors to descendants, much like genetic inheritance. They are subject to changes, similar to genetic mutations, through commonplace interactions between individuals, such as teaching and imitation (Manrubia and Zanette 2002). There have been a growing number of studies that attempt to apply the quantitative methods used by evolutionary biologists to the development and evolution of cultural traits (see, for example, Cavalli-Sforza *et al.* (1982)). Many traits studied, such as political preference, are subject to fashion and will change frequently over a relatively short time period, making them inappropriate markers for population and, more specifically, genetic structure (Manrubia and Zanette 2002). Surnames, however, represent cultural traits the transmission of which has clear parallels with genetic features. They are inherited from parents in the same way as the Y-chromosome (in the case of fathers to sons) or mitochondrial DNA (in the case of mothers to daughters)(Jobling 2001). In the case of the data used here, patrilineal (from fathers) inheritance dominates, meaning that much of the analysis is most applicable to the Y-chromosome.

George Darwin, son of Charles Darwin, initiated the use of surnames to investigate family lineage in 1875. He was interested in the frequency of first cousin marriages and whether their offspring experienced any adverse health effects as a result of this consanguinity (sharing an ancestor) (Darwin 1875). Darwin's and subsequent studies took marriages to be between a couple sharing an ancestor if they were isonymous. Isonymy, in this context, can be defined as the presence of identical surnames in the ancestry of a couple (Lasker 1968). A major breakthrough in the effective utilisation of surnames in genetics was made by Crow and Mange (1965) with their formalisation of the *Coefficient of Inbreeding from Isonymy* (Crow and Mange 1965) (see Equation 5.1). This measure, and those derived from it, forms the basis for many comparative studies of regions and their surnames (Colantonio *et al.* 2003).

Surname studies within genetics and more widely human biology are therefore based on the principle that to the extent that two individuals with the same surname are ultimately to share the same lineage, isonymy indicates biological relatedness (Lasker 1985). The hereditary nature of surnames and their tendency to remain highly concentrated in their areas of origin are the two traits most utilised. Hereditary surnames contain information about relatedness within populations because patrilineal surnames should be associated with a distinctive Y chromosome inherited from a male's father (Sykes and Irven 2000). However, the impracticality of collecting the genetic information of a complete population, past or present, makes proxy data, such as surnames, the only alternative in large-scale studies. Additionally, studies of extinct lineages have shown that many lines of descent quickly disappear so that the remaining individuals are much more likely to be related through a common ancestor (Lasker 2002).

2.3.1 ACCOUNTING FOR MULTIPLE LINEAGES WITH THE SAME SURNAME

The relationship between surnames and genetics is strongest where the founding population of an area was small, genetically diverse and each family group had a unique surname (Rogers 1991). Historical evidence suggests this is unlikely, as many founding populations were familial groups originating from the same region: such groups were likely to share a small gene pool and exhibit high levels of isonymy (Jobling 2001). Even with an "ideal" founding population there are some processes at work to cause deviations from equivalence between surnames and measurable relatedness. An example of this is genetic drift which is a stochastic process that pertains to the prevalence of genetic characteristics associated with particular groups or individuals. If, for one, or a variety of reasons, or simply by chance, an individual is more successful at passing on their genes (that is they have more children) than others, it will reduce the overall genetic diversity of a population. This impact of this will vary with the number of offspring the individual has and, more importantly, the size of the population: smaller populations will be more affected. A good example of genetic drift is the study of the medieval dynasty of *Uí Nèill* in Northern Ireland that

showed that 20% of all males sampled exhibited some kind of genetic relatedness (Moore *et al.* 2006). This is shown in Figure 2-2. It is also the case that a genetic lineage, as implied by surname commonality, can be broken through non-paternity (having a biological father different from who it is presumed to be), adoption and matrilineal surname transmission (King and Jobling 2009). It is possible to estimate non-paternity rates (thought to only be a few percent) in modern populations (see Macintyre and Sooman 1991) but is much harder to do for historical populations.

The final, and perhaps most significant, consideration is the fact that some surnames are inappropriate for implying a direct linkage to genetics. In many cases there are likely to be multiple founders of a single surname; the most obvious example being occupational surnames such as "Smith" (Hey 2000). It therefore follows that such surnames will have multiple lineages when compared with surnames acquired by a single individual or group in a specific area (King and Jobling 2009). This issue is one

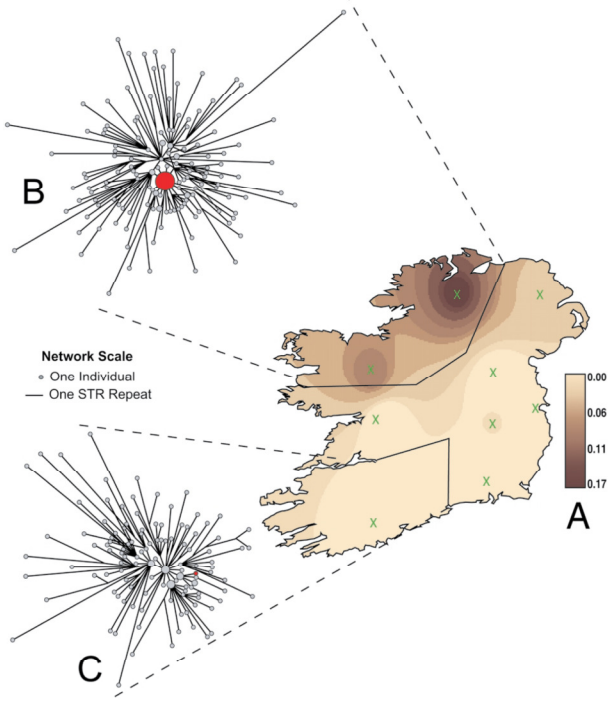


Figure 2-2: Phylogeography of the imh y-chromosome lineage associated with the Uí Nèill dynasty. (A) shows a clear concentration to the North of Ireland. A study of individuals removed from the area of highest concentration reveals that many shared a genetic trait (shown in red, (B)) that is virtually absent in the south of the country (shown in (C)). Taken from Moore *et al.* (2005: 2).

of the biggest motivations for the methods outlined in Chapter 4. The geography of individual surnames, even with contemporary data, gives a strong indication of their historical origins.

In summary, Jobling (2001) suggests the following 3 criteria for sound linkage between surnames and genetics.

1. The surname must have a unique origin.
2. There must be no illegitimacy that would introduce chromosomes from other surname groups.
3. Chromosomes associated with different surnames must have been unrelated at the time of surname establishment.

2.3.2 VALUE OF SURNAMES AND GEOGRAPHY IN THE CONTEXT OF GENETICS

Even in light of the significant associations demonstrated above, it is clear that surnames do not provide a direct substitute for genetic information. It should be emphasised that the purpose of using surnames to discern cultural and genetic structure within and between populations is to unearth general trends, not specific traits. Limitations surrounding the development of specific theories about individuals based on their surnames will undermine the wealth of useful information at a more general level. It is tenuous to suggest that all individuals sharing a surname are genetically related. Extending such conclusions to the female population would be even more inaccurate, given the wide international application of male-oriented surname conventions.

Careful thought is therefore required before surnames can be used in this context. That said, as will be outlined below and has been alluded to above, numerous studies have found compelling evidence of commonalities in the geography of surnames and certain genetic characteristics. Such studies result from either in-depth analysis of individual surnames or more general regional analysis. In both cases, surnames provide some key benefits. Firstly, it is relatively straightforward to obtain temporal

data, such as the 1881 Census used here. It would, by contrast, be impossible to undertake widespread genetic sampling of deceased individuals over a number of generations. Secondly, surname data has a much larger coverage and is more complete (in terms of representation of the population) than any genetics database is likely to be for the foreseeable future. For this reason there is a certain synergy between the two types of data. Surname data typically offers population coverage but only limited variability (typically name, frequency and location) whereas geneticists are more familiar with very small sample sizes but a very large number of variables for each. This latter point is especially relevant when attempting to characterise genetic traits associated with particular surnames.

The most compelling studies to provide a link between surnames and genetic traits have entailed ascertaining the origins of people in parts of the British Isles. Figure 2-3 is taken from Bowden *et al.*'s (2008) investigation into the genetic legacy left by Viking settlers in North West England. Each colour represents a proportion of a measurable genetic trait, known as a haplogroup, for populations randomly sampled

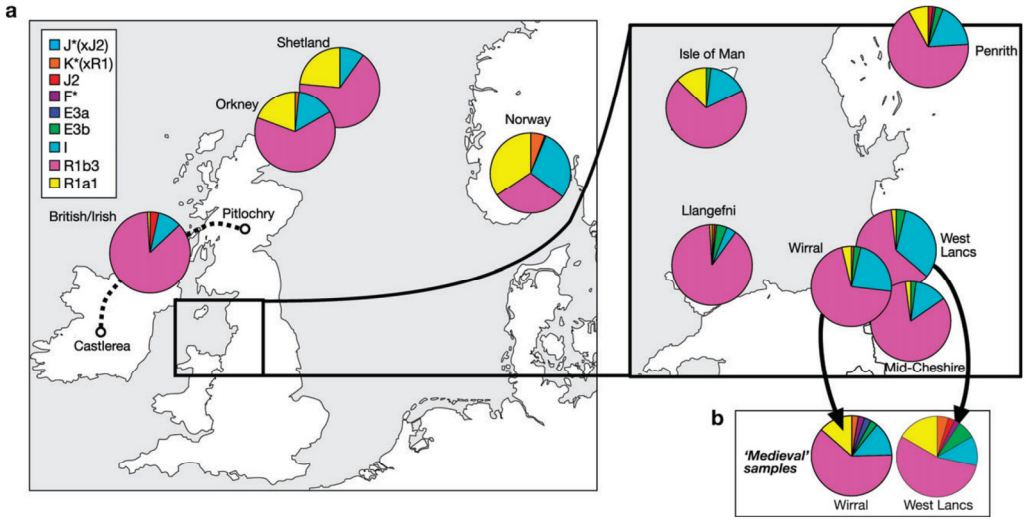


Figure 2-3: (A) The proportion of a randomly sampled population with particular genetic traits. There are no discernable differences between the Wirral, West Lancs and Mid-Cheshire charts and they bear little resemblence (in terms of proportion of R1A1) to Nordic populations. (B) The same traits measured in a population of people with known "Viking" surnames recorded in the medieval period. There are clear differences between these (in terms of R1A1 proportions) and the randomly sampled equivalents. In addition similarity with Nordic groups has increased. Taken from Bowden *et al.* (2007: 305).

in each area (a) and those sampled by surname (b). The region is widely known to have had a Viking presence and the results show that, by sampling modern individuals who bear surnames present during the Medieval Period, it is clear that they have a greater proportion of Scandinavian ancestry than the rest of the population (Bowden *et al.* 2007). Strong differences have also been found in Ireland between those with Irish surnames and those with surnames from elsewhere in the British Isles (Hill *et al.* 2000). The persistence of these distinctions suggests minimal interaction between groups over time and that, despite the mixing effects of modern migration, surnames continue to be an indicator of this trend.

At the European scale, numerous studies identify a strong geographic component to genetic structure across Europe (Novembre *et al.* 2008). There is wide agreement that this relates, at least to some extent, to the effect of distance and physical geographical features (Rosser *et al.* 2000). Such differences are likely to reflect the earliest settlers in Europe who clearly pre-date any linkages with surnames. There does, however, appear to be some debate about the extent to which more subtle changes in genetic structure are consequences of cultural barriers- such as language- or the continued influence of the physical constraint of distance (contrast Barbujani and Sokal (1990) to Novembre *et al.* 2008). In addition, Rosser *et al.* (2000), who do not believe language is a key determinant of genetic structure, concede that it is often the combination of physical and cultural factors that result in the strongest barriers to gene flow. It is therefore clear that some information is contained within cultural traits, such as surnames, especially when assessing the degree to which historical migrations (since surname conception) have disrupted the clinal transitions articulated in Rosser *et al.* (2000) and Novembre *et al.* (2008). The term "clinal" is widely used in population genetics and refers to gradual changes in genetic characteristics that occur over large areas.

At the regional scale, there are demonstrable links between surnames and genetics, but such links are- based on current research- less evident on the continental scale. As with all spatial studies, this appears to relate to the frame of reference of the research undertaken. Sub-national studies have used an intensive sampling framework designed to seek out the minor changes in genetic structure; this contrasts

with international studies that have sought much larger differences with relatively few samples. It is anticipated that, as genetic studies in the latter context become more detailed, subtler variations may become obvious, reflecting cultural traits, such as surnames, at finer levels of granularity.

2.4 CONSIDERATIONS IN THE ANALYSIS OF SPATIALLY REFERENCED POPULATION DATASETS

This chapter is concerned with a number of the distinguishing features and additional considerations required for the analysis of spatially- referenced population data. Whilst not unique to spatial analysis, these features have been overlooked in the context of surnames research. Datasets without a spatial component are less subject to variations in scale, or the methods of aggregation used to create them. To the initiated these are central considerations, but for those outside the geographical disciplines they are often unnoticed potentially leading to spurious results. This is not to suggest that this has been the case for all previous surnames research. Instead, it is to offer some clear context from the perspective of quantitative geography on the potential impact that spatial data's "special features" may have. This chapter outlines a number of key conceptual and philosophical issues associated with the analysis of spatial data and demonstrates their relevance to the spatial analysis of surnames. It is purposefully generalised as it provides the disciplinary context to many of the methodological decisions and analytical interpretations that follow in later chapters.

2.4.1 THEORETICAL FOUNDATIONS

The handling of spatial data and its analysis is rooted in quantitative geography, which emerged as a sub-discipline following the "Quantitative Revolution" of the early 1960s. The methodological and research outcomes from this "re-tooling" of geography have been much maligned, often unfairly (see Wrigley and Bennett (1981) and Fotheringham (2006)), by more recent geographic research, yet they underpin the extremely productive fields of spatial analysis and Geographic Information Science (GISc). Much of the early research in quantitative geography "borrowed" methods from other disciplines, primarily statistics, and attempted to apply them to spatial data (Openshaw 1984). Perhaps the biggest shortcoming of much of the pioneering research was its emphasis on techniques to the point where they became the focus at the expense of the phenomena being studied. As a consequence, some

have argued that many research outcomes were limited to little more than the identification of empirical regularities in spatial forms and movement patterns (Taylor and Johnston 1995). In addition, with the benefit of hindsight, the population data available and the technology to process it was, in many cases, inadequate to capture the phenomena of interest. For this reason, many relied on inference to stretch the insights provided by the new methodologies beyond acceptable bounds. The result was an artificial sense of objectivity reliant on induction and inference (Wrigley and Bennett 1981). An acknowledgement of the limitations of the approach taken with quantitative methods, particularly in the context of population research, has led to more useful applications and it is in this vein that the research for this thesis is conducted.

Quantitative analysis represents the only practical way of reducing larger datasets to a smaller amount of more meaningful information (Fotheringham 2006); the technological advances of the past two decades have facilitated this and dramatically widened the horizons of spatial analysis methods as a result (Taylor and Johnston 1995). Increased computing power has facilitated the use of quantitative analysis by a wide range of researchers and on datasets large enough to be considered truly representative of their target populations. National-level area classifications can now be produced using desktop computing infrastructure (see, for example, Vickers and Rees (2007)). That said, statements referring to an unprecedented volume of high quality spatial data can be found as far back as Haggett (1965) (and probably before). On the basis that the data and technology available in the 1970s was by modern standards inadequate, some question whether sufficient data can ever be collected to represent entirely the phenomenon of interest. Clearly the data have to sit within broader analytical methods and these are likely to provide important constraints. Spatial representation is largely dependent on the scale of the analysis and the spatial units used. These features are responsible for the "special nature" of spatial data (Haining 2009) and outlined in more detail below.

2.4.2 NATURE OF SPATIAL DATA

2.4.2.1 Scale

Scale is a fundamental consideration in geographical research and has multiple meanings. Longley *et al.* (2011b) provide the simplest definitions of scale as pertaining to the detail provided by the analysis, the extent of the analysis, or the level of cartographic representation on a map. Johnston *et al.* (2005) offer more complex definitions that refer to cartographic, methodological and geographical scale in the following terms: cartographic scale refers to the level of abstraction at which a map is constructed, whilst methodological scale refers to the spatial extent of the study undertaken. In the case of cartographic scale, the interpretation of "large-scale" and "small-scale" is the opposite to the other uses of the term, in which the former means highly detailed while the latter more generalized (Longley *et al.* 2011b). For clarity, scale is therefore not referred to in a cartographic context in this thesis.

Cartographic and methodological scales are under the control of the researcher, but geographic scale is more complex because it is inherent in many of the spatial processes themselves (Cliff and Ord 1981). This is important because it relates to the extent of the system of study that may change over time, with the locality of study and the resolution at which it is viewed. On this basis it is possible to create a nested hierarchy of levels at which the scale of a system can operate. Not all systems can be easily characterised in this way but it provides a robust framework for their study (Johnston *et al.* 2005). Such a hierarchy may exist in the "surname system", with a relatively static picture of broad linguistic distinctions appearing increasingly subtle if viewed at the European scale and more fluid, distinctions appearing if viewed at a national scale and much more dynamic transitions occurring if studying surnames locally. To establish if this is the case, three levels of the hierarchy (Continental, National and Local) are considered in the analyses of Chapters 5 and 6. Clearly the magnitude of differences in surname composition is scale-dependent in the sense of geographic extent. When clustering into 5 groups the between group difference is likely to be much less on a local scale than on a continental scale. It is therefore not possible to make direct comparisons between two classifications in terms of the

degree of dissimilarity if their scales are inconsistent. Put another way, the magnitude of difference required to allocate data to different groups is likely to be much smaller on the local scale when compared with the national or continental scales.

The conceptualisation of scale in this way is yet to take place in the surname literature because surnames have not been viewed in the context of a series of nested systems with different levels of interaction. The standard approach has been to utilise whatever datasets are available and stretch the inferences as far as possible with few, if any, linkages made between local studies and those at the national and continental levels. The creation of such a hierarchy would better inform the applications of surnames research in other fields and provides a meaningful framework for study.

2.4.2.2 Spatial units

The configuration, in terms of size and shape, of the spatial units analysed may potentially have a major impact on the research outcomes (Openshaw 1984). Even if data are representative of individuals, much of spatial analysis is reliant on comparative measures, such as rates or ratios, which must be calculated through inference or using aggregate units. To this end the majority of, if not all, studies treat space as being formed of either discrete objects or continuous fields. Such terms were introduced into the GIScience literature in the late 1980s and early 1990s, and have since come to dominate thinking about human conceptualisations of geographic space (Goodchild *et al.* 2007). The best form of representation is subject to debate and may largely be applications dependent. In the case of population data, individuals are discrete, autonomous units in space but it is rarely the case that data are available at this level. Individuals are therefore subject to some form of aggregation into larger, often administrative, spatial units or the cells of a grid.

Aggregation, whatever form it takes, has the effect of smoothing out the variation in the data (Haining 2009). Clearly this smoothing can ultimately result in over-simplification and excessive information loss. It therefore follows that smaller areal aggregates are preferable to larger ones because they are likely to be relatively

homogenous and, as size reduces, more representative of their populations. The drawback, however, is the offsetting of greater spatial precision by a reduction in statistical precision (Haining 2009). As spatial units represent fewer people, so they are more likely to reflect random variation in the data, especially when calculating rates or ratios; this is often referred to as the "small numbers problem". There is therefore a trade-off between the increased homogeneity of smaller spatial units, but greater noise, and greater statistical precision at the loss of internal heterogeneity. A loss of internal heterogeneity may in fact be a preferable outcome in a number of contexts, especially when painting a subtler picture of more general spatial patterns (Dorling 1995).

Aside from the amount of aggregation commonly associated with the size of the spatial unit (in terms of the number of people it represents), its shape is also important. When assigning population to grid cells (the standard continuous field representation), it is common to use squares (they can be any regular shape that tessellates) orientated vertically in the field of view. Aggregation to irregular polygons, such as administrative boundaries, is much more complex because there is an almost infinite combination of both shape and size that could be used (Openshaw 1984). Real world geometrical arrangement is a product of physical constraints, such as political boundaries, or can be determined by a population threshold to ensure a roughly consistent level of population aggregation (see for example Martin (2002)).

A useful aspect of gridded units and surface models, however, is the way in which they facilitate direct comparison across multiple datasets through the simple aggregation or disaggregation of cells. This is particularly effective when undertaking a comparative study between inconsistent geographies (see Bracken and Martin (1995)). Often the most pragmatic solution is to attach attribute information to the centroids of the polygons and redistribute it to a gridded geography using one of the many areal interpolation methods available (such as Goodchild *et al.* (1993), Bracken and Martin (1995)). Surface models of population can be also be used as a basis for discretised geography through the identification of boundaries and thresholds based on the aggregation of grid cells according to similar attributes, such as socioeconomic

characteristics (see Martin 1998a). As will be seen in Chapter 4, this characteristic is particularly useful when specifying thresholds in surname density surfaces.

Aggregation, facilitated by changing the spatial unit geometries, potentially leads to the Modifiable Areal Unit Problem (MAUP) (Openshaw 1984). The MAUP is well-known and comprises of both scale and aggregation effects. The former refers to the increased strength of correlation between two variables with increasingly aggregate units; whilst the latter is demonstrated by the fact that changing the configuration of the spatial units (whilst keeping the underlying population distribution fixed) will produce different correlation coefficients (Johnston *et al.* 2005). Research into the effects of the MAUP therefore cautions against the use of a single level of aggregation for data analysis. This advice has been overlooked by the surnames literature with all published research dependent on a single type of, generally administrative, spatial units and a preference for relatively few of them when establishing the strength of correlations between distance and surname compositions (see for example Scapoli *et al.* (2007)). In this thesis, where possible, multiple aggregations are used to test the consistency of the results at a variety of spatial scales.

Despite these criticisms it can nevertheless be argued that it is acceptable to use some types of generic or "imposed" spatial units in the analysis of population data because they bear some correspondence with the underlying spatial distribution (Openshaw 1984). On the continental scale, for example, many uncontested national borders are drawn along cultural/ linguistic lines and therefore represent a logical aggregation when investigating such things. At a finer level of granularity it is often the case that historical boundaries, such as parishes in Great Britain, are not imposed because they were often drawn between distinct communities and are seen by many as a sensible aggregation of historical populations (Pooley and Turnbull 1998). In addition, more recent attempts have been made to create administrative spatial units that are more representative (both socially and in terms of density) of underlying population structure. Martin (1998b), for example, describes the automated creation of census geography (in this case Census Output Areas (OAs) for Great Britain) with a requirement that they exhibit a degree of internal homogeneity with regard to social

composition. In this example social composition is measured according to proportions of households falling into different tenure classes; demonstrating that increasingly automated zone design can begin to account for demographic factors (Martin 1998b).

Efforts to include social characteristics aim to reduce the uncertainty in spatial research by attempting to create natural rather than imposed units of analysis appropriate for their application(s) (Martin 1999). It is possible that aggregations of surnames can represent such a unit of analysis in a number of contexts. The insights they provide are sufficiently generalised so as to be applicable to the majority of the population within each region but they are also sufficiently informative so as to provide logical regions for their use in population research. It is conceded that the regions themselves are inductive aggregations of previously aggregated data (because their basic building blocks are not individuals). However, surname geographies (as will be demonstrated in later sections and is discussed above) can correspond to the administrative units that have developed in order to cater for the needs of the specific communities in which they are found.

2.4.3 CONCEPTUALISATIONS OF SURNAME GEOGRAPHY

Like many other studies of population attributes, there is yet to be a universally agreed method of spatially partitioning surnames. The methods used to partition space in surname research are especially important as the results are often transferred into other fields (such as genetics) with different conceptions of space. If, for example, the spatial distribution(s) of surnames are to be compared with data on genetic structure, similar units of analysis are required to enable direct comparison. In the case of genetic transitions, they are treated as "clinal" or continuous but punctuated by more abrupt "barriers". Barriers are physical or cultural features, such as mountains or water bodies, which may prevent two or more population groups from interacting. Linguistic comparisons take a more discrete view, with analysis focussed on the mapping of uniform regions produced using clustering. This has become the dominant approach and provides the classification of areas based on

aggregate measures of surname diversity, there has been relatively little consideration of how best to handle the spatial distribution of individual surnames.

As was suggested in Section 2.2.1, the study of individual surnames has moved little beyond visually interpreting point distributions (or counts) of surname instances. Kaplan and Lasker's (1983) attempt to address this limitation provided some important insights into the regional dynamics of surnames and demonstrated the idea that frequencies often follow standard models of distance decay. On this basis, a continuous field view of surname distributions would be meaningful. The implicit assumption in the research, however, is that - at least for the surnames studied - there is a single core area of concentration that becomes increasingly diffuse with distance from the centre. In reality this is not the case: many surnames have multiple points of origin and their distributions are altered by settlement geography. Manni *et al.* (2005) have undertaken a more advanced analysis with Dutch surnames through the use of self- organising maps (SOMs). For 9000 of the most popular surnames in the Netherlands they identified areas of highest concentration and probable origin. Whilst providing a discrete view of surname distributions, and therefore a view less compatible with many studies of genetics, the research is a significant advance, especially in its analysis of thousands of surnames individually. The work presented in Chapter 4 attempts to combine a more continuous interpretation of surname distributions with the practical advantages of discretisation and provides a number of additional metrics associated with surname distributions unobtainable with Manni *et al.*'s (2005) approach.

Studies that aggregate multiple surnames conceptualise geography as either continuous or discrete. Sokal *et al.* (1992) take a continuous approach (outlined above) in line with studies of genetics and produce frequency surfaces of 100 surnames that are then combined to find common boundaries (dramatic changes in the frequency distribution). This study attempted direct comparisons between genetics and surnames by suggesting that abrupt surname boundaries were the result of barriers to population movement and mixing. Interestingly the researchers found no such relationship, instead suggesting that the boundaries were the product of historical factors related to the origin of surnames (Sokal *et al.* 1992). Nonetheless,

stronger associations between surname transitions and physical barriers to gene flow have since been found in other studies beyond Britain through the use of barrier algorithms (see Chapter 5), such as Monmonier's algorithm (see Manni *et al.* 2004). The most compelling is the implementation of the algorithm in the Ferrara Province of Italy, where a number of the identified barriers closely match topographic features known to have restricted population movement. The approach, however, has only been published in a few studies (from the same authors (Manni and Barrai 2001, Manni *et al.* 2004, 2008)) and would therefore benefit from further research.

The use of discrete spatial units to create uniform surname regions is another common approach and the one that informs Chapters 5 and 6. The creation of uniform groups (the number of which is pre-determined by the researcher) is common within several disciplines and is informed by a long tradition of classification. Within surname analysis this method has been used to group areas based on the similarity of their surname compositions at European level (Scapoli *et al.* 2007) through to small-island level (Branco and Mota-Vieira (2004)). Classification approaches, based on dissimilarity matrices, are less easily translated onto the concepts of isolation by distance and, as Chapters 5 and 6 demonstrate, have a number of limitations relating to the requirement for a fixed number of groups. There are, however, a number of methods available to mitigate these limitations. Such methods are yet to be fully considered in the literature and their inclusion in this thesis marks another improvement to current research.

2.4.4 WIDER DISCIPLINARY CONTEXT

The title of this thesis refers to the concept of population structure, which is a commonly used term in a variety of disciplines. In the context of population genetics, for example, it represents the geographic limits placed on relatedness. Barriers to migration, cultural differences or simply isolation by distance create these limits and are also manifest in clear transitions between groups. The configuration of the similarities/ differences is referred to as the structure. In demography, the spatial component of population structure is of less overall importance, attention is given to

attributes such as the proportions of different age groups, genders and levels of education. As Chapter 4 outlines, a clear spatial structure was established during the process of surname creation and acquisition and this spatial structure is likely to be manifest in other socio-economic, cultural and genetic structures. The purpose here is to establish the extent to which such characteristics persist in contemporary populations.

Several fields of population geography and demographics are concerned with population structure and one of the most relevant here is geodemographics, or area classification. Geodemographics is the analysis of people by where they live (Sleight 1997) and is premised on the idea that the location of an individual or group of similar individuals provides useful demographic information (Harris *et al.* 2005). Geodemographics in this broad sense is relevant for a number of reasons. Firstly, this thesis can be thought of as producing a geodemographic classification based on what is revealed by the spatial distributions of people's surnames. Chapter 7 demonstrates how surnames can be used, for example, to improve sample designs in population genetics through targeting specific areas of surnames. As will become clear, this, in principle, is little different from a retailer or health care provider, informed by geodemographics, targeting a specific area based on its socio-economic characteristics.

A second important commonality between geodeomographics and this thesis is a shared interest in classification and clustering. The clustering of large population datasets has become routine in area classification (Adnan *et al.* 2010), but there has been relatively little research into improving the algorithms used to create the classifications. Similar algorithms are used in a variety of fields, not least population genetics. In this case the datasets are characterised by much smaller sample sizes (in terms of number of individuals) and thousands more variables. Large numbers of variables increase the potential for instability in the clustering - a problem well researched in this field (see Simpson *et al.* (2010) and Shimodaira (2004)). In this thesis, the more robust clustering methods from genetics will be applied to the exceptionally comprehensive surname data, guided by the geodemographic practices of classifying areas.

Previous surnames research has sought to represent similarity or differences between contiguous regions largely in the context of clinal, or gradual, changes in the distributions punctuated by the occasional abrupt transition. There has been relatively little interest in the apparently close linkages, in terms of surname compositions, between regions that are geographically distant. Only one study, Longley *et al.* (2007), has sought to establish any causality and impacts of these linkages. Such regions are often quite small and therefore likely to have been missed by the relatively coarse granularity of previous research. The idea that similarity/ difference is more than the product of geographical distance has long been recognised in geodemographics and represents important exceptions to the idea that populations in close geographical proximity are likely to be more similar than those further apart (based on the often quoted "Tobler's First Law of Geography" (Tobler 1970)). For example, Webber and Longley (2003), in their discussion of geodemographics, highlight the importance of social similarity that is independent of locational proximity. Whilst location is important, they argue that its impact varies depending on the attributes measured; for example there are greater similarities in family composition within a neighbourhood when compared to the age of the residents (Webber and Longley 2003). In light of this, as will be discussed in Chapter 6, contiguity constraints in the construction of classifications have been discarded.

Like geodemographic classifications, this thesis provides an inductive representation of similarities or differences between population groups based on common population attributes. The results of this process can provide useful context for future hypothesis generation and the study of population characteristics beyond a direct relationship with surnames. The potential of the insights provided by both the methods and data outlined below go beyond the level offered by standard geodemographic classifications. They represent an analysis of ubiquitous phenomena that appears relatively stable over time. The data utilised are impartial and can be analysed at a variety of scales to produce meaningful conclusions that may capture both historical and contemporary trends. In addition, it is hoped that the methods utilised below can be easily replicated to produce consistent results for different

datasets. All that is required is a location, its surnames and their respective frequencies.

2.5 RESEARCH NEEDS

This chapter has described why the spatial distribution of surnames is neither uniform nor random and how their spatial patterning makes surnames an important indicator of population structure. The study of surnames as a spatial data source (and the resulting applications) is in its infancy and would benefit from further research. Based on the literature reviewed above, this penultimate section will outline some key research requirements that need to be met if the use of surnames as a credible source of quantitative data is to expand.

Population geneticists were the first to see the value in surname data and they are largely responsible for their continued analysis. A major shortcoming resulting from the limited interest of geographers and spatial scientists is the lack of consideration for uniquely spatial problems, such as the impacts of scale and size of spatial unit in the results (Openshaw 1984). This is especially evident in the regionalisation of surnames. There have been no studies, for example, that vary the size of the populations or number of spatial units in the analysis. This is partly due to the poor availability of detailed spatial data associated with surnames and also an apparent lack of awareness of such issues. This thesis seeks to address this shortcoming by placing many of the methodological foundations developed in population genetics in the context of 50 years of quantitative geography and spatial analysis research. It is hoped that the results will serve to increase the credibility of surname analysis within spatial disciplines and provide a basis for future research.

As Point 1 of Jobling's (2001) criteria (Section 2.3.1) demonstrates, geography should play a key role not only in the selection of a surname in the study of genetics but also the spread and interaction of that name over time. The purpose of surnames in this context is to act as a filter through which to select members of the population or specific regions. Filtering individuals by surname has been less researched due to the previous complexities and volume of data required to identify surnames that have a single lineage. Only one paper (Manni *et al.* 2005) has sought to identify the origins (or areas of highest concentration) of a large enough volume of surnames to facilitate their use in the large-scale genetic sampling of an area. Previous attempts to select

surnames have been manual and informed by the largely descriptive genealogical literature. There is therefore a need for a comprehensive and geographically extensive database that identifies probable areas of origin, current areas of concentration and diffusion over time for as many surnames as possible. Chapter 4 outlines how this has been achieved using a number of established spatial analysis techniques.

The compelling link between surnames and genetic structure makes almost certain that other population characteristics can be discerned from such analysis. Very few studies have attempted to do this. For example, Piazza *et al.* (1987) demonstrated in *Nature* the possibility of inferring migration rates from surname data but this paper's citations do not expand beyond the early population genetics literature cited above. The limited use of such applications is also clear from Darlu *et al.* (2011), who modelled migrations between regions using Bayesian statistics and logistic regression. One of the key reasons for a lack of migration studies is the need for roughly comparable surname data for two time periods. As a greater number of historical records, such as the 1881 Census, become digitally available such studies should increase. In addition, only one study (Longley *et al.* 2007) has attempted to combine the historical insights provided by surnames with more contemporary demographic indicators. This research illustrates many possible lines of productive research for demographers and geographers that are yet to be followed.

2.6 CONCLUSIONS

The purpose of this chapter has been to suggest that surnames are representative of cultural, linguistic and genetic phenomena at a range of geographic scales. These associations are manifested in the clear geography of surname frequency distributions (that has, in many cases, persisted over generations). This is attributable to both the inheritability of surnames and the relatively static nature of populations. Unfortunately, due to in part to a lack of data, there has been relatively little interest in the spatial analysis of surnames. Instead, much of the research has focussed on establishing the degree to which surnames can be treated as a proxy for genetic information. Such studies have failed to consider many of the common issues, such as scale and choice of spatial units, tackled over 50 years of quantitative geography and the resulting spatial analysis techniques. The outcome has been a dependence on over-aggregated spatial datasets. As will be demonstrated in subsequent chapters, the quality of the data utilised here enables a thorough consideration of these issues and more robust outcomes as a result.

It is acknowledged that surnames do not provide perfect correlates between culture and genetics but that they nevertheless provide ubiquitous phenomena and are one of the few easily measurable universal cultural attributes. The data reflecting this, such as telephone directories and censuses, are available in increasingly convenient formats to facilitate analysis on an unprecedented scale. The purpose of this thesis, in the light of what has been outlined above, is to provide a sound methodological and analytical framework for the spatial analysis of surnames by utilising the most comprehensive data available.

3 SURNAME DATA AND PRELIMINARY ANALYSIS

The process of spatial analysis often follows a number of well-defined steps. These include problem formulation; obtaining the data; exploratory analysis; hypothesis formulation; appropriate analysis and testing; and, finally, review (de Smith *et al.* 2009). This chapter is concerned with steps two and three above: the acquisition and appraisal of data and their preliminary analysis. To this end it is divided into two clear sections. The first provides an overview of the three surname data sources used here. It includes information about their quality, spatial referencing and how representative they are likely to be of the populations concerned. The second section undertakes some preliminary analysis of the data for Great Britain. Its purpose is to offer insights into the spatial distributions of surnames at both individual and aggregate levels. Such insights provide the context to many of the processes captured in the more substantive analysis undertaken in Chapters 4 to 6. They also serve to enforce and advance many of the results outlined in the previous chapter from other studies investigating the spatial distributions of surnames.

3.1 DATA

The spatial extent, granularity and number of records represented in the data used in this thesis are unprecedented in the context of surname analysis. They have been obtained from three sources: 1881 Census of Population for Great Britain, an enhanced version of the 2001 Electoral Register for Great Britain and the UCL Worldnames database (worldnames.publicprofiler.org). The latter contains surname frequency data for almost 30 countries (including 16 from Europe), from these sources it has been possible to analyse up to 220,000 spatial units of population within Great Britain and to cluster data pertaining to over 152 million individuals at the European level. The volume and quality of the data analysed far exceeds previous studies of surnames and population structure (see Scapoli *et al.* (2007) for comparison).

3.1.1 1881 CENSUS OF GREAT BRITAIN

Used in this thesis are the returns from the 1881 Census for England, Scotland and Wales. The data provide the names and place of enumeration (Parish and Registration District) for 29 million people, with a total of 425,000 unique surnames (approximately 346,000 of which have occurrences of more than 10 people). The frequency distribution of the top 500 surnames in the 1881 Census is shown in Figure 3-1. When digitising the Census records, volunteers from the Church of the Latter Day Saints reproduced surnames exactly as transcribed in the original with the following exceptions: double barrelled names had dashes removed; spellings with unusual punctuation were excluded; spaces in Mc and Mac names have been removed; and those names only surviving as initials or only containing two letters were also removed (Barker *et al.* 2007). The data and their metadata are available from the UK Data Archive (see Wooland and Allen 1999).

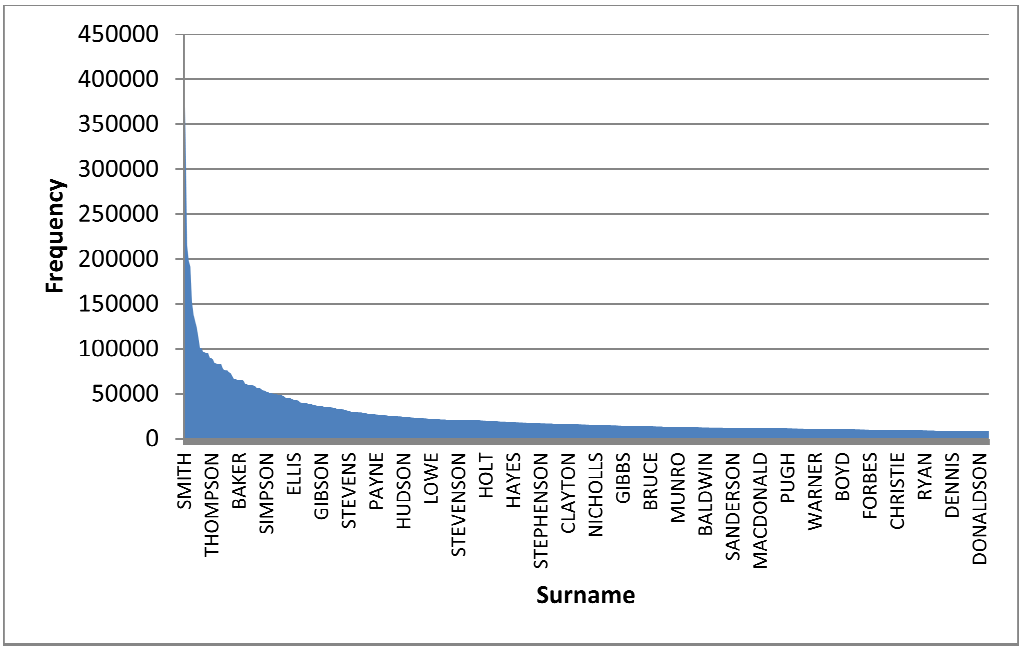


Figure 3-1: A plot showing the population of each surname (X axis) against the top 500 surnames in Britain for 1881 (Y axis). Even within the top 500 surnames a long tailed distribution emerges. Only a selection of surnames are labelled.

The geography of the 1881 Census is complex because of ambiguity surrounding some of the administrative boundaries used. Indeed, the census report states that the boundaries used "overlap and intersect each other with such complexity that enumerators and local registrars in a vast number of cases failed altogether to unravel their intricacy" (Woolland and Allen 1999: 49). From the available boundaries, it was thought sensible to use Registration Districts, as opposed to Parishes or Counties. Registration Districts are much less coarse than counties but coarser than Parishes and provide the best balance between spatial resolution and a sufficient population size to obtain a representative population of surnames within each geographical unit of analysis. Analysing Registration Districts also makes pragmatic sense as their boundaries have been digitized and are available for download from the UK Borders website (http://edina.ac.uk/ukborders/).

In this study, 662 Registration Districts were mapped, of which 658 have surname data, with the remaining classified as common land or missing data. These latter districts were removed by enlarging the neighbouring districts contiguous with them.

The average population in each district is approximately 4,900 inhabitants: Figure 3-2 shows a population density map in 1881 built from Registration Districts.

The 1881 Census is likely to have a number of sources of error and uncertainty. Human error would have been introduced at a number of stages, from when the surname was first recorded on the day of the census through to its much more recent digitisation. The errors from the former are likely to be far greater than any modern census because the enumerators were required to interpret verbal information from a largely illiterate population (Barker *et al.* 2007). In addition to the spelling and recording of the correct surname, the assignment of people to Registration Districts has also been problematic. Registration Districts (used here) are not perfect aggregations of contiguous 1881 Parishes (the smallest geocoded unit). In a number

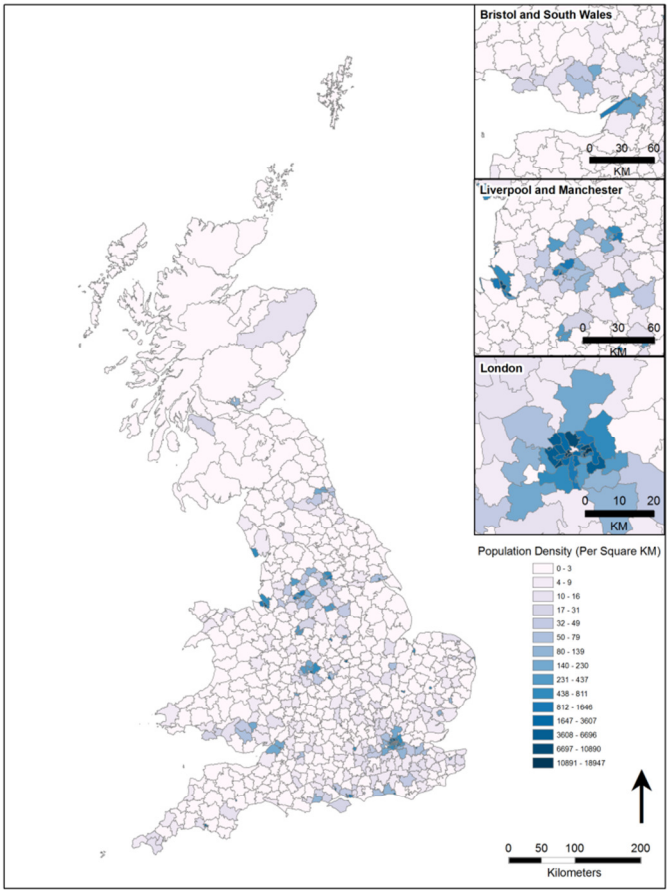


Figure 3-2: A map showing the population density of each 1881 Census Registration District. As can be seen most districts had a low population density, with only a few urban districts possessing high populations.

of cases a Parish would straddle two Registration Districts. In these instances the entire population of the offending Parish may or may not have been assigned to both Registration Districts. This inconsistent policy reflects the considerable confusion, both on the day of the Census and its subsequent digitisation, surrounding the 1881 enumeration geography (Woolland and Allen 1999).

It is acknowledged that the limitations of the 1881 Census are much greater than the 2001 Electoral Register used for comparison. It is thought, however, that the data quality is adequate in the context of much of this thesis' generalized analysis.

3.1.2 2001 ENHANCED ELECTORAL REGISTER OF GREAT BRITAIN

The contemporary surname frequencies for Great Britain are taken from an enhanced version of the 2001 Electoral Register. It solicits the names and addresses of all residents aged 16 or over (irrespective of eligibility to vote in UK or EU elections), with a view to compiling a list of eligible voters during the period of currency of the Register and eliminating addresses with no eligible residents from follow up enquiries. The full version is made available to government and credit reference agencies only, under strict terms of confidentiality. A public version is made available for sale to marketing and other organisations, and comprises electors who have not 'opted out' of inclusion in the full version. Using data from a range of commercial sources, this version is enhanced by marketing organisations, in this case CACI (UK) Ltd., in order to provide a more complete coverage of the population. There is likely to be bias against inclusion in the Register for members of ethnic minorities and privacy sensitive socioeconomic groups. These are known to concentrate disproportionately in particular areas, particularly within conurbations: the modified register is nonetheless the most comprehensive source of individual names data available for the UK. The data record 41.6 million people resident in Britain in October 2001 (the Census of 2001 records 59.8 million people of all ages), possessing a total of 828,131 surnames (of which 711,000 represent fewer than 10 people): this represents approximately 73% of the total population recorded in the 2001 Census of Population, with most of the remainder comprising minors – the

omission of which is unlikely to contribute bias in data use to represent the geographic distribution of surnames. The frequency distribution of the top 500 surnames is shown in Figure 3-3. No version of the Electoral Register has been made publicly available for Northern Ireland for many years, for reasons dating back to heightened security concerns during 'the Troubles[3]'. For this reason the analysis focuses upon Great Britain and not the entire United Kingdom.

The 2001 enhanced Electoral Register holds data at the level of an individual's address. This facilitates any level of aggregation either within the existing geographies available or according to bespoke units. Many of the standard spatial units used for population data in Britain are based on the 2001 Census of Population geography using the National Statistics Postcode Directory (NSPD) (available from http://www.ons.gov.uk/). From each unit postcode, the data may be aggregated into any of the following available 2001 Census administrative boundaries (sequenced

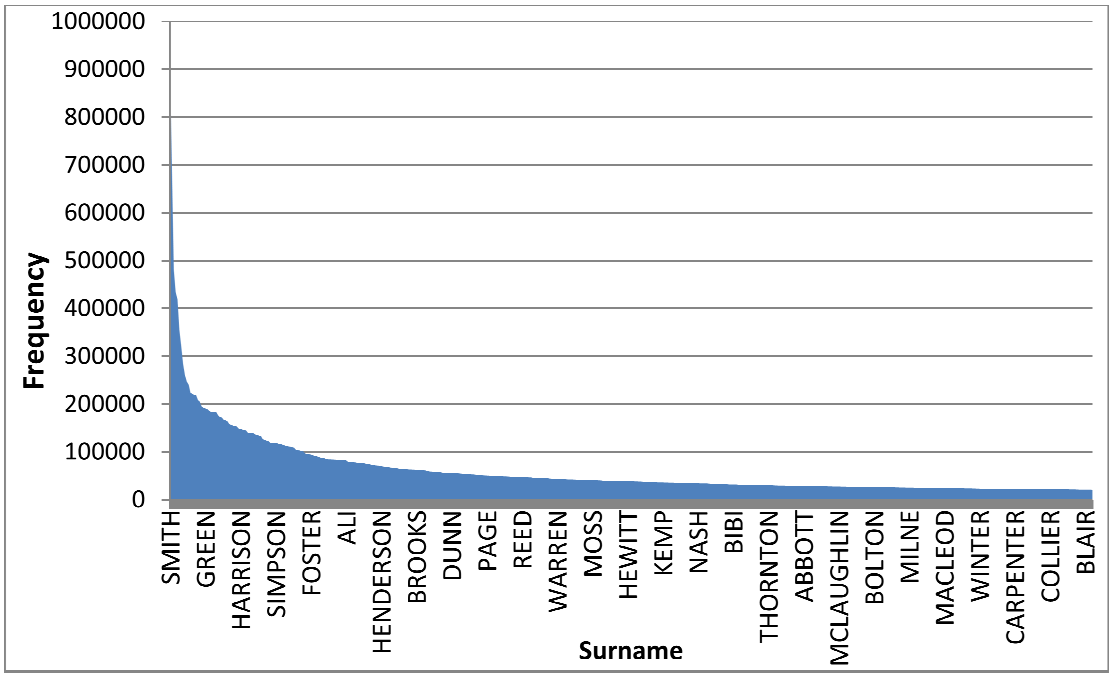


Figure 3-3: A plot showing the population of each surname (X axis) against the top 500 surnames in Britain for 2001 (Y axis). Even within the top 500 surnames a long tailed distribution emerges. Only a selection of surnames are labelled.

[3] A period of ethnoreligious- political conflict in Northern Ireland considered by many to have ended with the "Good Friday" Agreement of 1998.

here from smaller to larger areas): Output Area (OA), Lower Super Output Area (LSOA), Middle Super Output Area (MSOA), Super Output area (SOA), Local Authority District, or Government Office Region (GOR). In addition there are a range of non-census geographies available in Great Britain that have been based on other established processes such as elections in the case of the c. 10,500 Wards or historical patterns of land ownership in the case of Civil Parishes. Figure 3-4 shows the population density at Local Authority District level.

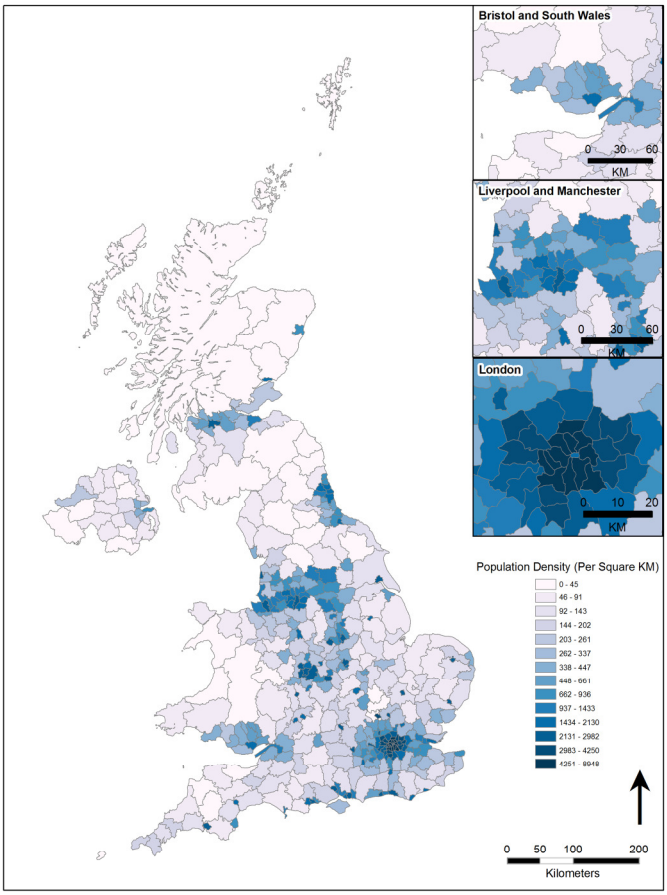


Figure 3-4: A map showing the population density from the 2001 enhanced Electoral Register of each Local Authority District. These are designed to contain approximately the same number of people. Urban districts are therefore smaller in area and have a higher population density as a result.

3.1.3 EUROPEAN SURNAME DATA

The UCL Worldnames database, containing the names and addresses of in excess of 400 million people in 26 countries, is derived from publicly available population registers and telephone directories collected during the 2000-2005 period. From this, data from 16 European countries are extracted (including Great Britain), comprising over 8 million unique surnames, their geographical locations and their frequency. A list of countries, name frequencies and geographical characteristics is shown in Table 3-1. Their raw population density is mapped in Figure 3-5. Due to the wide range of data sources used here, this map reflects- in part- data density.

The motivation to focus on Europe, as opposed to all countries in the database, is twofold. Firstly, as outlined in Chapter 2, it has been subject to the most research.

Table 3-1: The countries and their data used in this study. "NUTS Level" refers to the geographic unit of analysis used.

Country	Data Year	Total Population	Worldnames Population	No. Unique Surnames	NUTS Level
Austria	1996	8,316,487	2,520,012	81,387	2
Belgium	2007	10,511,382	3,489,068	852,492	3
Denmark	2006	5,457,415	3,074,871	153,134	2
France	2006	64,102,140	20,280,551	1,197,684	3
Germany	2006	82,314,900	28,541,078	1,226,841	2
Great Britain	2001	60,587,300	41,690,258	828,131	3
Italy	2006	59,131,282	15,927,926	1,305,554	3
Luxemburg	2006	480,222	117,619	75,267	3
Netherlands	2006	16,570,613	4,672,344	531,970	2
Norway	2006	4,770,000	3,536,524	123,240	3
Poland	2007	38,518,241	8,015,455	339,339	2
Rep. of Ireland	2007	4,239,848	2,916,744	46,507	3
Spain	2004	45,116,894	9,545,104	260,469	3
Serbia, Montenegro and Kosovo	2006	10,159,046	1,704,559	69,977	2
Sweden	2004	9,142,817	791,143	135,830	3
Switzerland	2006	7,508,700	1,565,098	19,270	3
	Totals	426,927,287	148,388,352	7,247,092	

Secondly, and more pragmatically, the countries chosen represent the largest group of neighbours in the database. Those included therefore reflect the available data with omissions reflecting an inability to obtain the requisite data; the sourcing of data is ongoing. The amount of work required to source, clean and geocode the surnames in this dataset should not be underestimated and has required a large amount of expertise and computing power to achieve (see Adnan 2011).

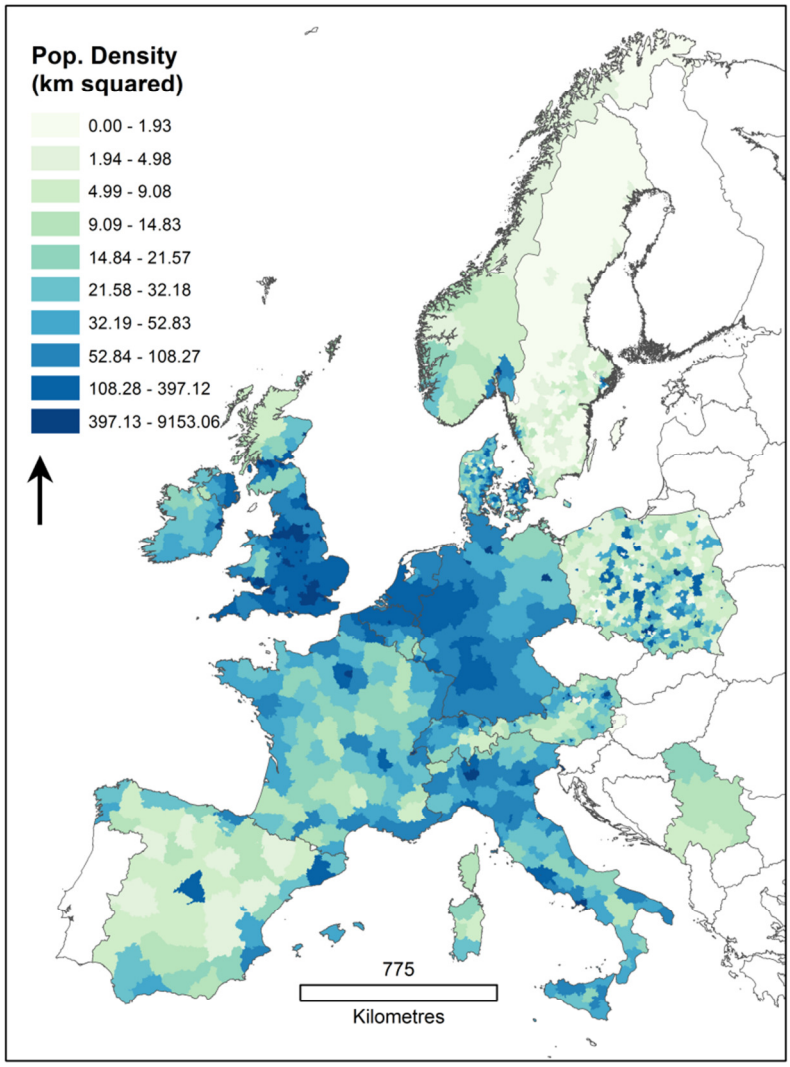


Figure 3-5: Population density for the 16 countries used here. This is based on the available data and is therefore reflective of the number of individuals in the database rather than actual population as the counts have not been grossed to reflect national populations. The spatial units are the mix of NUTS 2 and NUTS 3 used in the analysis.

Unlike the Electoral Register for Great Britain, the Worldnames database does not comprise data from a single source. Instead it has been compiled from a range of providers with different levels of disclosure and geographic referencing. As such, while most of the entries are at individual address level, there are some cases, such as the data for the Republic of Ireland, where only aggregate-level name frequencies are known. Individual addresses have been carefully geocoded to a set of geographical coordinates (latitude and longitude) at levels of resolution ranging from the building level to the administrative region, going through street name, postcode, city and metropolitan area. Since the concern here is with general, continental-level patterns, ideally the detailed locations will be aggregated onto a set of standard geographical regions of similar size and population. Across the European Union (EU) the NUTS regions (*Nomenclature d'Unités Territoriales Statistiques*) provide a convenient set of geographic units that broadly conform to these aims. NUTS are a standard referencing system for the hierarchy of five levels of administrative sub-divisions of EU countries for statistical purposes, from broad country regions (NUTS 1) to municipalities (NUTS 5). Initially all surname data were aggregated to NUTS 3 level (the province or department), however it became apparent that some countries with a relatively large number of NUTS 3 units (such as Germany where these correspond to 429 *Kreise* or Districts) relative to their population were having an undue influence on the results (outlined in Chapter 6). Therefore, for this thesis a combination of NUTS 2 and NUTS 3 regions have been chosen in an attempt to address this problem and produce a set of homogeneous areas in terms of population size and geographical extent. Table 3-1 lists for each country the NUTS level selected. This resulted in a total number of 763 geographic units across the 16 countries.

3.1.4 TREATMENT OF RARE SURNAMES

Many very low-frequency surnames, with less than 10 occurrences for example, are likely to arise from slight differences in the spelling of more common surnames, or errors in the recording process. To account for these in an automated way is extremely complicated and often inaccurate (see, for example, Snae (2007)) so many previous studies have resorted to manual selection and merging of surnames based on the researcher's knowledge. One of the key contributions of this research, however, is to unearth patterns through the use of automated, repeatable,

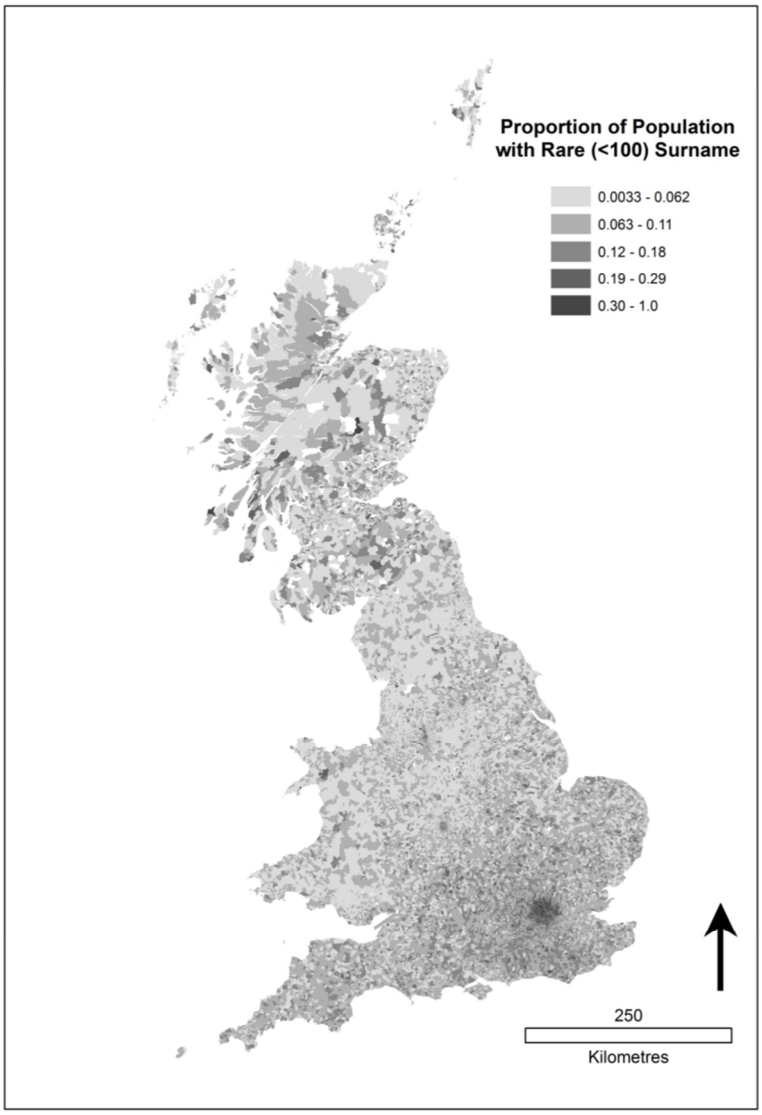


Figure 3-6: A map showing the proportion of the British population with a rare surname according to the 2001 Enhanced Electoral Register. In this case rare is defined as a frequency < 100 and mapped at OA level.

computational methods capable of handling the millions of records. A manual approach is therefore not appropriate. It is also the case that the definition of a rare surname can depend on application. For example, for the purposes of seeking to establish core areas of a surname's concentration, the topic of Chapter 4, it was considered necessary to remove the large number of "rare" names to fulfil the criteria for sound statistical inference. For reasons explained in the chapter it was thought best to focus on the 30,000 surnames with 100 or more occurrences.

In contrast to Chapter 4, the approach taken to regionalise surnames in Chapter 5 includes all the surnames recorded in the data. As is shown by Figure 3-6, at an aggregate level, there is a clear geography to rare surnames: in Britain at least, which may indicate distinctive population characteristics. As the preliminary analysis of surname distributions shows, there is only minor variation in the impact of such surnames between spatial units, even at fine scales.

3.2 PRELIMINARY ANALYSIS

The following section is devoted to preliminary analysis of the data for Great Britain outlined in the previous section. The purpose here is to demonstrate that surnames exhibit a clear spatial structure that has endured over several generations. The first aspect will show some simple spatial distributions of single surnames before forming a more aggregate picture through a variety of measures. The final section here combines the statistical insights provided by power-laws to map changes in the frequency distributions of surnames between 1881 and 2001.

3.2.1 SURNAME GEOGRAPHY: SOME EXAMPLES

Figure 3-7 shows the uneven distribution of a selection of surnames when looking at their density across a west-east transect of Great Britain. Each of the 9 surnames has been classified as Welsh (Wel), English (Eng), Cornish (Cor) or Scottish (Sct) reflecting their known places of origin (according to Hey 2000). It is clear that even with contemporary data there is not a uniform distribution. Trescothick, for example is still found only to the far west of the country while all the Welsh surnames selected show abrupt reductions at the approximate Easting of the English-Welsh border. The Scottish names are perhaps the most uniform with two distinct peaks that correspond to the population centres of Glasgow and Edinburgh. Mapping the relative frequencies of surnames using location quotients (see Equation 4.1), as has been done in Figure 3-8, tells a similar story: even popular surnames, such as Richards, are most prevalent at or near their geographic areas of origin.

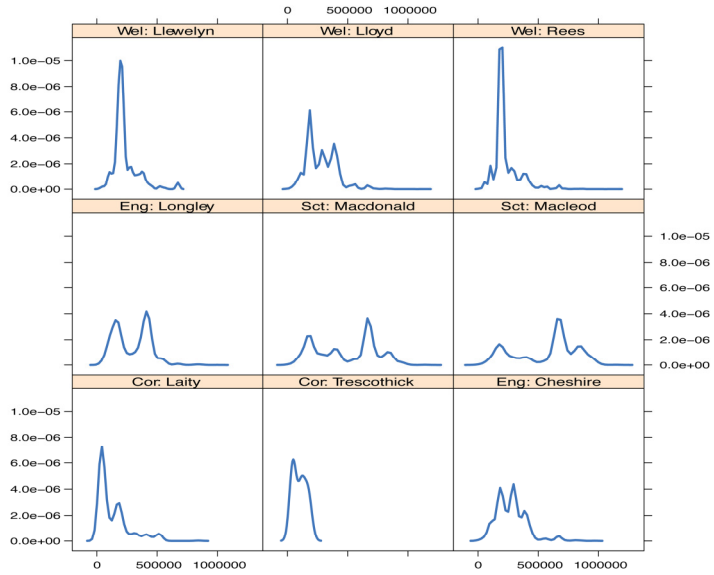


Figure 3-7: Density plots (based on raw counts) of nine surnames plotted against easting (British National Grid). A selection of English (Eng), Scottish (Sct), Welsh (Wel) and Cornish (Cor) names have been selected to demonstrate characteristic geographical distributions of such names. Published in Cheshire *et al.* (2010).

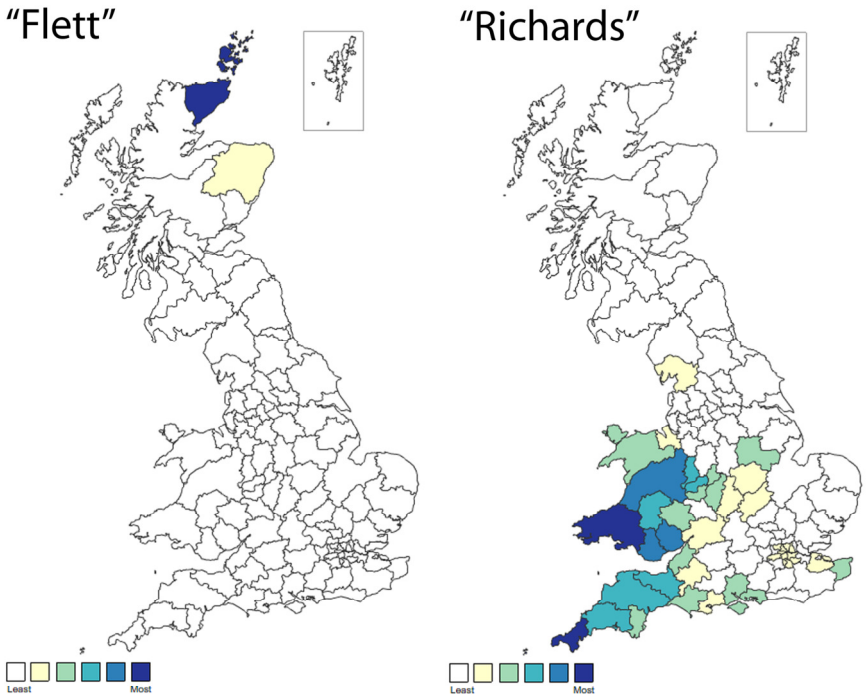


Figure 3-8: 2001 distributions of the surnames "Flett" (left) and "Richards" (Right). Mapped using the location quotient (Equation 4.1). Source: gbnames.publicprofiler.org.

3.2.2 SURNAME DIVERSITY

It is also interesting to see the relationship between variations in the abundance of surnames relative to population size. The use of the Gastner and Newman (2004) method to create a cartogram using the populations of the c.220, 000 OAs as a basis to alter the size of the spatial units is shown in Figure 3-9. The purpose here is to reflect the number of people each OA represents thus emphasising the uneven distribution of the British population and the associated relationship between highly populated areas and their surname compositions. Figure 3-9 illustrates the diverse surname compositions of Great British cities in contrast with rural Wales and the Scottish Islands, which have the smallest number of surnames per person. A closer look at this detailed map does, however, reveal a number of exceptions to this rural/ urban distinction in surname diversity. Parts of East London, Tower Hamlets for example, have very low surname diversities because of the large Bangladeshi community with characteristically few surnames. Such anomalies occur across a range of surname diversity measures (see McElduff *et al.* 2008). Figure 3-9 also illustrates the importance of the impact of urban areas on Britain's contemporary surname geography. As the preferred locations for both national and international migrants urban areas may, in some cases, require special treatment when trying to discern the underlying population structure shown in British surnames for the purposes of genetic sampling.

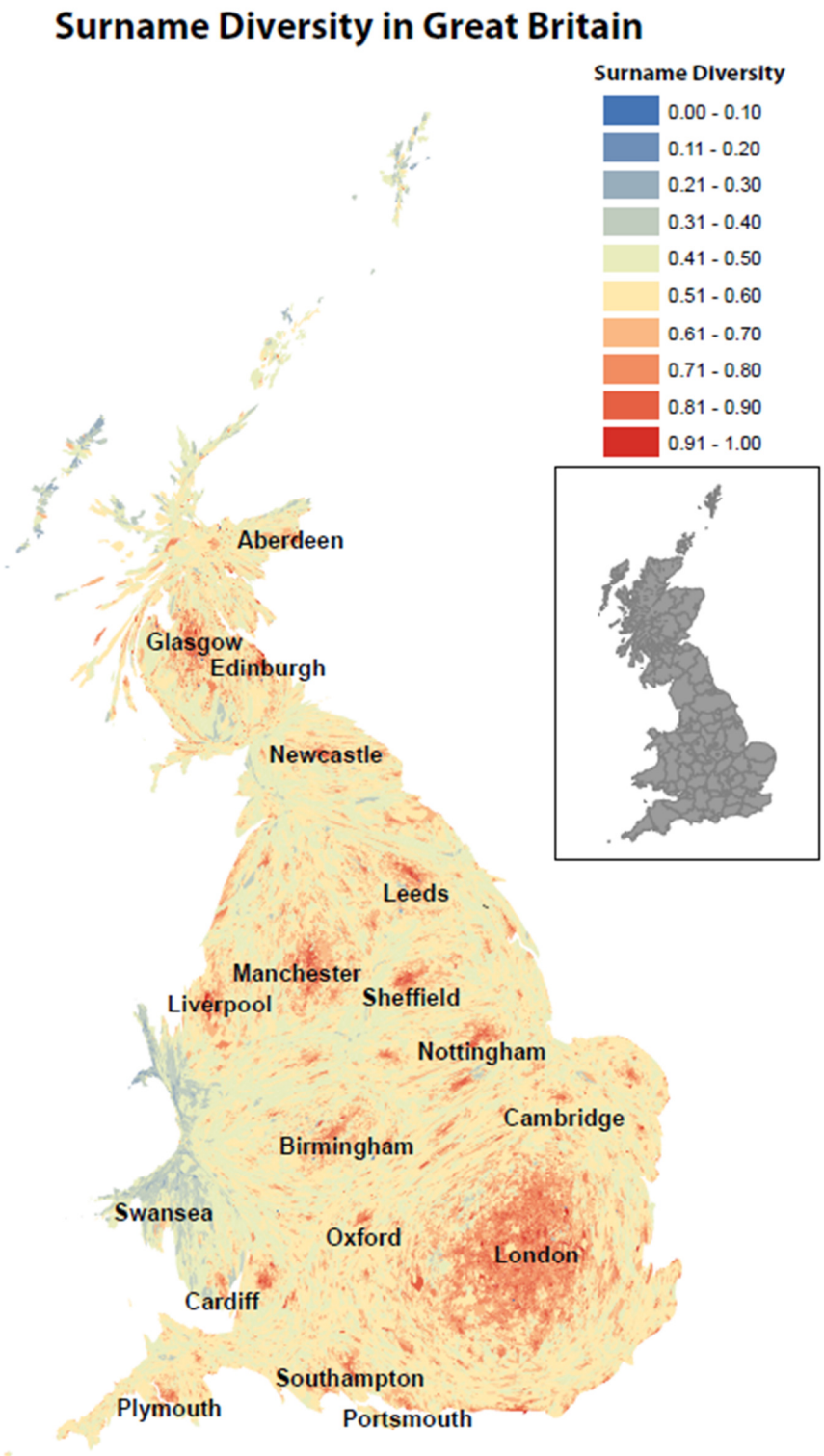


Figure 3-9: A cartogram showing the diversity of surnames (number of surnames divided by population) at Output Area (OA) level in Great Britain. The OAs have been scaled by their population size and clearly illustrate the high diversity of surnames in cities as compared with more rural areas. Wales has particularly low surname diversity.

3.2.3 SURNAMES AND POWER LAWS

The relative number of surnames per head in an area does not say anything about the frequency distribution of surnames. For example, if a population of 100 has 0.5 surnames per head there is no indication as to whether every surname is shared by precisely two people or 50 people all share a single surname and the remaining 50 each have a unique surname. Based on Figures 3-1 and 3-3, it is clear that British surname frequencies, like those of many other countries, are characterised by very long-tailed distributions. It therefore follows that the majority of surnames are rare, but that the majority of people do not possess a rare surname. This results in an extremely uneven distribution of surnames amongst the population that can be captured using a power-law (Manrubia and Zannette 2002). It is also striking that the pattern is also seen when subdividing the population into all but the smallest geographical scales (from, for example national-level to small administrative district level) (Fox and Lasker 1983). This means that although the surnames themselves may change over space and time, the typical proportions of people with the most popular surname (whatever that may be) to the least popular surname remain relatively unchanged. The completeness of the dataset is an important feature in determining power laws. Incomplete data can create misleading impressions, regardless of whether or not the data can be summarised with a power law (see Bentley *et al.* 2011). The enhanced version of the Electoral Register for Great Britain is the most complete individual-level representation of the British population available and so provides the best representation of surname frequencies in this context.

Figure 3-10 was produced by grouping surnames from the 2001 Electoral Register (triangles) and 1881 Census (circles) by their numbers of bearers (the surname 'population') on the x-axis and counting the frequency with which each 'population' size occurs on the y-axis. The plots are less intuitive than those produced in Figures 3-1 and 3-3 but they reveal more information. Points towards the top left represent the large number of times a rare surname occurs and the final point to the bottom right is for the most popular surname ("Smith") with a very high population that is not exceeded by any other surnames. In the context of the many events over the course of the 20th Century (such the two World Wars and the continuing influx of

migrants) capable of having an impact of changing the population's surname distribution, the two plots are surprisingly similar. The surnames they represent have changed (there are many more in 2001) and the increased frequency values of the 2001 points reflect the larger population, but the gradients of the lines are almost identical. It also provides reassurance that, despite not being conducted to "modern" standards and with the introduction of additional errors in the digitisation process, the 1881 Census has captured the likely distribution of surnames present at the time. Had there been major errors then it is likely that the distribution in Figure 3-10 would have deviated from the power law.

Figure 3-10 suggests a stable distribution of surname frequencies across the population at the national level over the course of the last century. This global impression, however, may mask many local transitions in population structure. In an

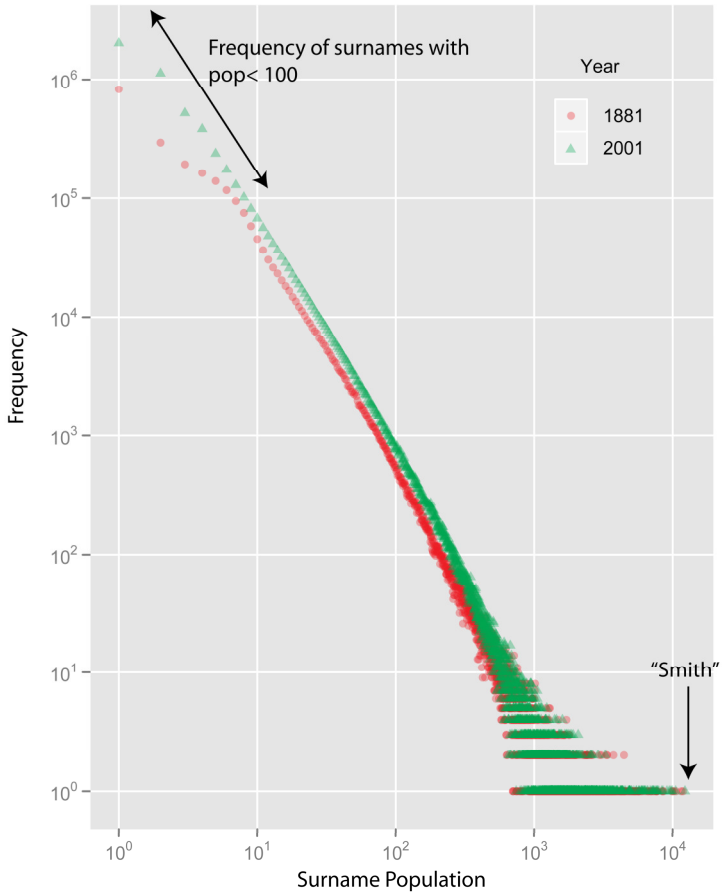


Figure 3-10: A plot showing the power law relationship of the number of times (frequency) a surname population occurs. In this case the population of the surname Smith (in excess of 900,000 for 2001) occurs only once but lesser population sizes of less common surnames occur many times.

attempt to capture these, the alpha value (gradient) of a line fitted to the surname population distribution (using the same types of plots as Figure 3-10) of each 1881 Registration District was calculated for both 1881 and 2001 data. The 2001 surname counts were aggregated from individual level to 1881 spatial units, thus enabling a direct comparison between the two years. A number of papers (see Fox and Lasker (1983), Panaratos (1989) and Manrubia and Zannette (2002)) have developed models for the analysis of surname power-law distributions. In this case the line fitted is calculated using the *plfit* tool developed by Clauset *et al.* (2009). Here the widely used tool is treated as a "black box" and therefore not subject to further discussion.

The gradient of the power law is more informative than simply looking at how many surnames per-head of population there are (as shown in Figure 3-9). It gives a sense of the distribution of surnames amongst the population of interest. As Figure 3-11 illustrates, shallower gradients (lower alpha values) suggest a dominance of common surnames compared with rare ones and vice versa with steeper gradients. One would expect the former for Wales, known for its relatively small number of surnames, and the latter for urban areas. Figure 3-12A demonstrates that this, for the most part,

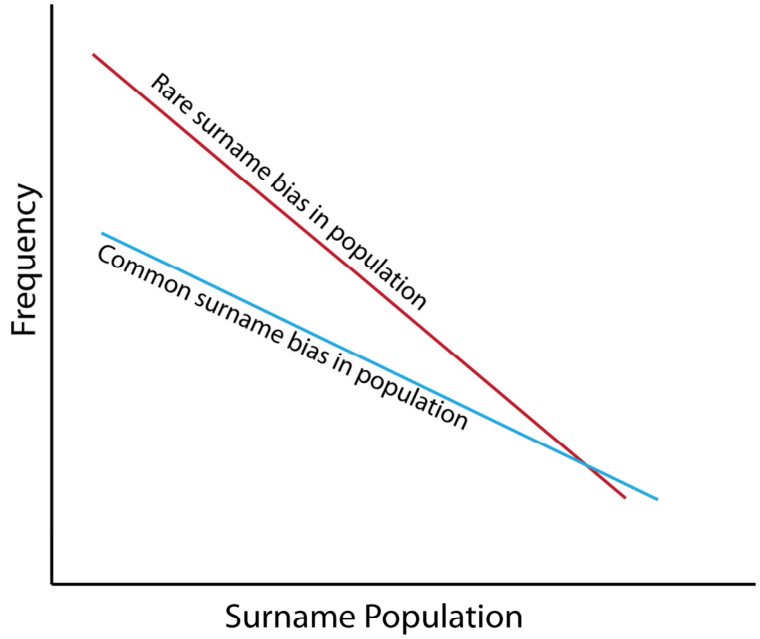


Figure 3-11: Demonstration of common gradients associated with lines fitted to populations who tend to possess more common surnames (in blue) and those who have a bias towards rare surnames (in red).

holds true with the 1881 data. In 2001 (Figure 3-12B) the lowest gradients, however, are found in Scotland and the Northwest. Of most interest is Figure 3-12C, which shows the change in gradient (2001 gradient subtracted from 1881 gradient) between the two time periods. This demonstrates that, for the majority of Great Britain, there has been a reduction in the number of surnames: proportionately more people have more common surnames. This is expected with fewer than half the surnames per head of population in 2001 versus 1881. The exceptions to this general rule are shown in yellow. The majority of such areas are either urban (including London, Southampton, Manchester and Liverpool) or in Wales and the far north of Scotland. The relative increase in power law gradient for urban areas is another example of the impact of, largely international, migrants but in rural areas may reflect even a small influx of domestic migrants or reduction in "native" population.

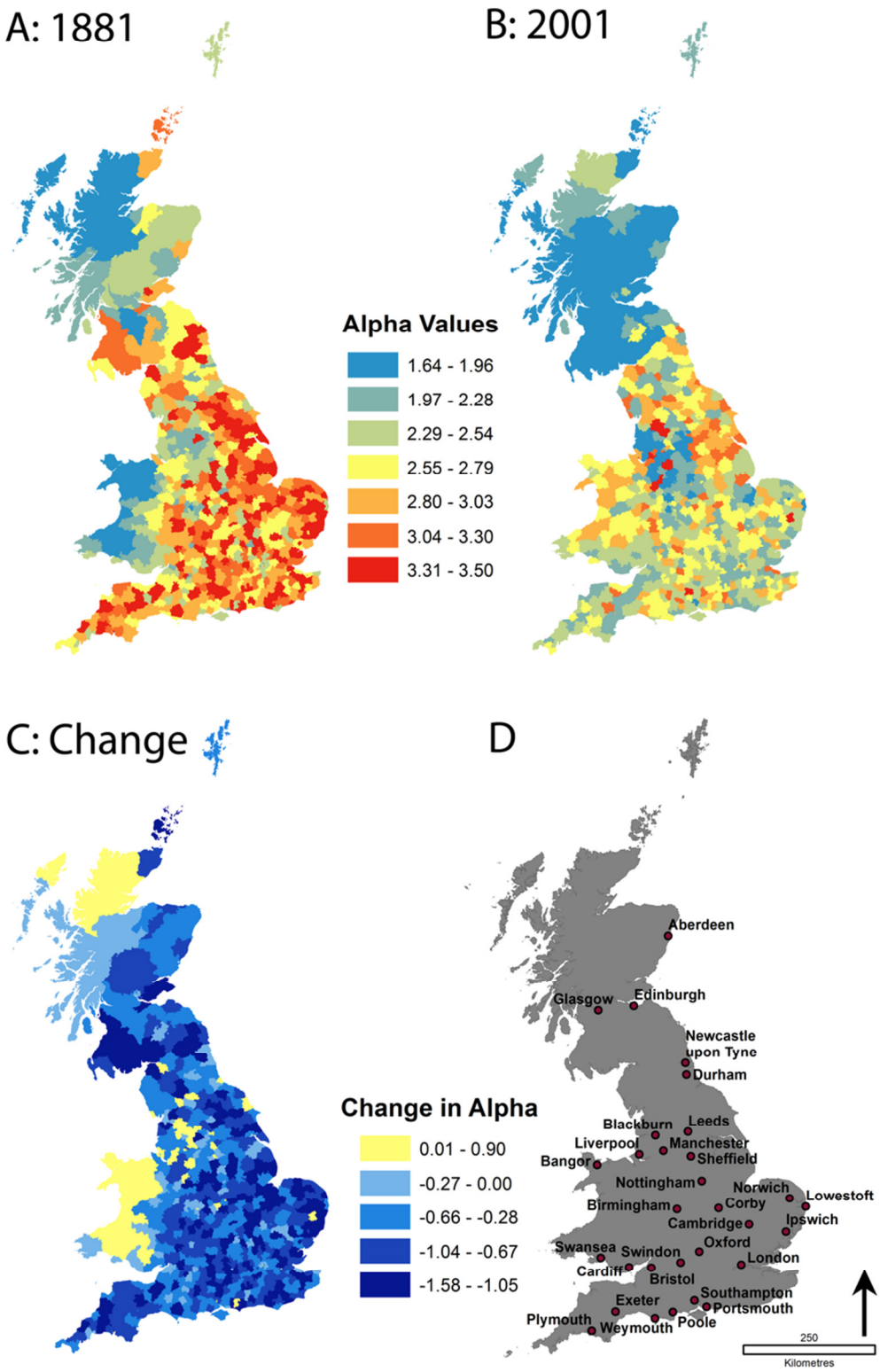


Figure 3-12: Maps of the respective power law gradients (alpha values) for 1881 (A) and 2001 (B). Change between the two years is shown in (C). (D) shows some larger towns/ cities for orientation purposes. 1881 Registration Districts are used for both maps.

3.3 CONCLUSIONS

This chapter has outlined the data used in this thesis. It is unrivalled in its comprehensiveness, in terms of both spatial granularity and representativeness of the populations in the 16 countries studied. The most comprehensive datasets, in the form of the 1881 Census and 2001 Electoral Roll, pertains to Great Britain and facilitates the most detailed analysis undertaken in this thesis. The scalability of the methods outlined in Chapter 5 will be tested using the more generalized data from the 16 European countries in the Worldnames database in Chapter 6.

It is important to conduct exploratory analysis of any data before utilizing increasingly complex methods to reach more substantive conclusions. The preliminary analysis conducted here substantiates the assertion in Chapter 2 that surnames are not geographically random phenomena. Instead they exhibited clear and persistent distributions over space and time that are likely to reflect the cultural context in which they were created. There is a clear geography to surname distributions, as each surname remains associated with its area(s) of origin in Great Britain. In addition, whilst the composition of surnames within each area varies spatially and temporarily, there is a surprising consistency, following a power-law, in the proportions of individuals associated with each surname, from the most common to most rare.

4 TOOLS TO DISCERN SPATIAL PATTERN: DETECTING SURNAME CLUSTERS

It is clear from the overview of surname research provided in Chapter 2 that there is as yet no systematic, comprehensive and automated method for discerning the spatial characteristics of individual surnames. The reliance on simplistic maps from the genealogical literature or studies based on partial datasets limits the potential for surnames to be used as a research tool. This chapter outlines the creation of a surname typology for Great Britain, using data from both 1881 and 2001, which contains a large number of metrics based on the identification of area(s) of highest concentration for individual surnames. This provides indicators not only of a surname's geographic origin in the country (self-evident only for toponymic names) but also of its current spatial extent and spatial relationship with other surnames and place names. In addition, temporal comparisons can be made that unearth population stability or movement. The results of this analysis are stored in a database that can be easily accessed by other researchers who wish to select surnames with particular spatial characteristics.

The need for a spatial typology of surnames is reflected in a growing number of papers, especially within population genetics, in which surnames have been selected according to their spatial patterns (Manni *et al.* 2005). One of the most straightforward applications is the identification of the approximate area(s) of origin for a particular surname. In addition, as is demonstrated in Section 4.3, it is possible to argue that analysis using historic and contemporary data can provide baseline and change measures, which are useful in studies of migration, amongst others.

As discussed in more detail in Chapter 2, surnames have diverse origins that can be characterised by unique geographic patterns. Previous research into the spatial distributions of single surnames has relied on little more than visual interpretations of

mapped surname distributions (see Mascie-Taylor and Lasker (1990), Porteous (1982), and Hey (2000)). Any interpretations drawn from these are subject to the known limitations of human perception of spatial clusters (see Rogerson (2006)). Despite an awareness of such shortcomings there is yet to be a consistent set of heuristics developed to characterise the spatial distribution of individual surnames. Manni *et al.* (2005) provides the closest example of this, but their approach with self-organising maps (SOMs) still requires manual disaggregation of patterns as part of the analysis. A key advantage to the appropriate application of well-tested spatial analysis techniques is the fact that many have been designed to remove such subjectivity. In these cases, interpretation will be left until after the analysis, when previous knowledge and experience can be applied to assess the plausibility of the outcome, and used to make comparisons between the different distributions in a robust and transparent manner (Rogerson and Yamada 2009). In addition, the use of computational spatial analysis, as opposed to manual methods, facilitates the automation of the methodology, rendering it applicable to tens of thousands of surnames.

Most previous analysis has also failed to consider fully the effects the underlying population density of a target area and its impact on the relative concentrations of surnames. Only Kaplan and Lasker (1983) explicitly account for variations in the underlying population by undertaking chi-square tests to assess variations in observed and expected numbers of surnames in urban areas. Areas of high population density are likely to have received the largest numbers of migrants of regional, national and international origin. It therefore follows that such areas will have an increased likelihood of occurrence of any particular surname (McElduff *et al.* 2008) and could skew the mapped distributions towards them.

In addition, this research is the first to consider individual international migrant surnames in the analysis. Such surnames identify areas that have been subject to changes in population structure, and that have often been destinations for different migrating groups. Urban areas, the boundaries of which are not always crisp and well defined, often fall into both of these categories: indicators of surname diversity and

extent might thus be used as criteria upon which to base comparisons between rural and urban areas, and to create more meaningful distinctions between the two.

The approach taken here treats the discovery of areas of highest concentration in a surname's distribution as a spatial clustering problem. The identification of geographic concentrations or clusters is well understood, and a number of robust solutions and associated statistics are available. What follows is an outline of four methods considered in this analysis alongside a number of preliminary results. This informed the decision to use kernel density estimation (KDE) in the final methodology. Details of the final methodological steps undertaken to create the typology are explained in detail before a selection of results are discussed. The final aspect of this chapter explores the temporal aspects of the analysis and how it can be used to chart historical population processes.

4.1 SPATIAL CLUSTERING

The detection of spatial clusters is a widely studied and commonly applied aspect of spatial analysis that is concerned with determining the degree of randomness in the distribution of spatial objects across an area (Longley *et al.* 2011b). A pattern can be random with its points located independently and all locations equally likely; clustered with some locations more likely than others and points attracting each other; or dispersed with the presence of one point making others in the vicinity less likely (Longley *et al.* 2011b). These three outcomes are the product of different interactions within the spatial data and, in the case of surnames, the former two are most likely.

There are many methods available for the identification of spatial relationships and clustering between data points. A number of these have been selected based on their potential to identify areas of clustering in individual surname distributions. The purpose here is to provide an overview of the methods trialled and some of their key outcomes.

As discussed in Section 2.4, spatial data can be conceived as discrete objects or as a continuous field. The investigation of spatial clustering conforms to these distinct views with the additional consideration that most continuous representations require discrete points as input. The two discrete methodologies will be considered first, before the two continuous approaches are outlined. The former (location quotient and Moran's I) have been designed for areal data while the latter (KDE and spatial discontinuities) use spatial points as input to create surface representations.

4.1.1 BRIEF NOTE ON DATA

Both the 1881 Census and 2001 Electoral Register (see Chapter 3 for more information) have been used for this analysis. The examples provided below are drawn from each with the majority of the preliminary analysis performed on the 1881 data at Registration District level. The more substantive methodological development

in Section 4.1.5 was undertaken with the 2001 data at OA level to ensure that the final approach taken could handle the high volume of data. Where point data is required the centroids of 1881 Registration Districts and 2001 OAs are used.

4.1.2 DISCRETE METHODS USING AREAL DATA

4.1.2.1 Location Quotient

The location quotient (LQ) is a straightforward measure that compares an area's share of a particular activity with the share of that activity at a more aggregate spatial level (Burt *et al.* 2009). In the present context it can be defined as follows:

$$LQ_i^j = \frac{A_i^j / \Sigma_{i=1}^n A_i^j}{B_i^j / \Sigma_{i=1}^n B_i^j} \tag{4.1}$$

where A_i^j is the frequency of surname *i* in spatial unit *j,* B_i^j is the frequency of surname *i* in Great Britain and *n* represents the total number of surnames. LQ values of greater than 1 identify spatial units with a higher concentration of a selected name than would be expected if the surname had a uniform distribution across Britain. A surname's core locality can be established using this measure by selecting the spatial units with LQ values that exceed a predetermined threshold at the top end of the distribution. This is likely to be subjective but can be informed by the distribution of LQ values and their pattern of contiguity across all spatial units.

If a surname has a discernable core area(s) of concentration it is likely that the spatial units with the highest LQ values will be contiguous or in close proximity to one another. There is a greater likelihood of contiguity with more aggregate spatial units, such as the 1881 Registration Districts on the basis that many of the fine scale variations will be smoothed over. If there are large distances (allowing for differences in the configuration of spatial units in rural versus urban areas) separating many of spatial units that share high LQ values, the distribution of the name is best described as dispersed. Figure 4-1 shows the spatial distributions of the LQ values for 6 illustrative surnames in 1881. The top three (Grahamslaw, Forster, Pedlar) are demonstrative of surnames with a discernable core region, with Grahamslaw being the most extreme example of this. The bottom three surnames (Smith, Gray and

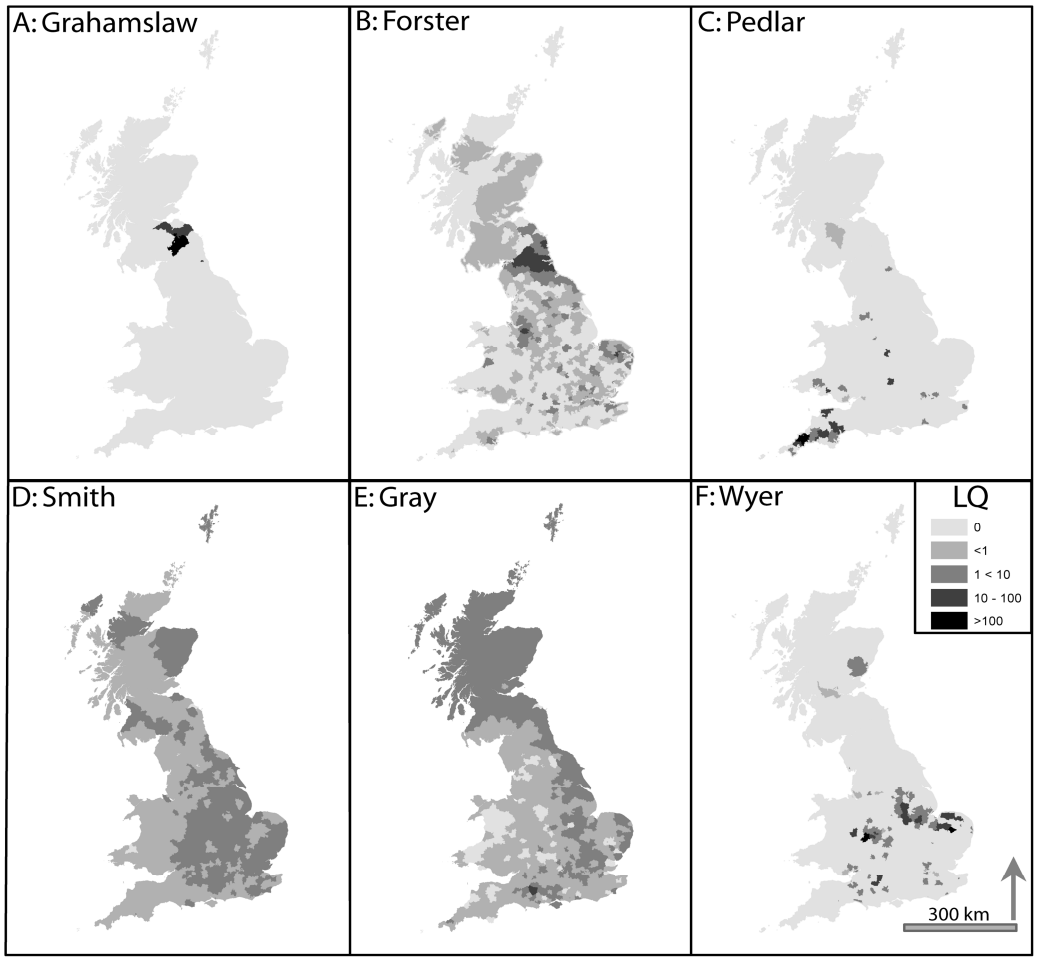


Figure 4-1: The spatial distributions of the LQ values for 6 surnames in 1881. Darker colours reflect higher LQ values and greater relative concentrations of a surname in a particular area. Published in Winney *et al.* (2010).

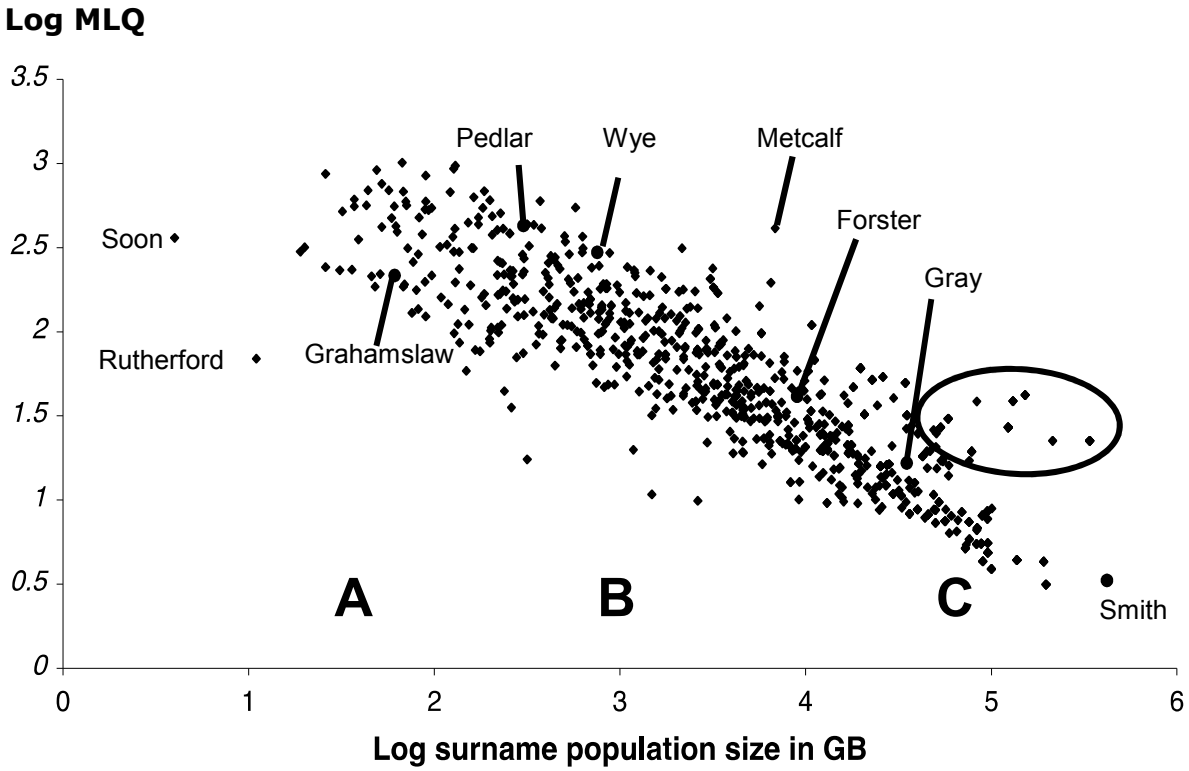


Figure 4-2: Graph of the log(MLQ) of the Registration District with the highest LQ for each surname (Y-axis) against Log(surname population size) in the 1881 Census (X-axis). There are a number of surnames (circled) with a higher MLQ than might be expected for the surname sample size (Jones, Davies, Evans, Thomas, Hughes, James and Phillips), which are established Welsh surnames. The surnames from Figure 4-1 are also marked. Published in Winney *et al.* (2010).

Wyer) are clearly more dispersed and characterized by much lower LQ values, and especially in the case of Smith and Gray, no areas of discernable clustering.

In addition to the spatial distribution of LQ values, the range in LQ values for a given name across Great Britain provides an indication of spatial clustering: a small range is indicative of an even spread across Britain, whereas a large range suggests that a small number of (contiguous or otherwise) spatial units account for a large proportion of a particular surname in a few areas with few, if any, other occurrences. Perhaps more interesting is the relationship between the surname's maximum LQ (MLQ) value and its frequency. The MLQ refers to the area where the surname is most concentrated in Great Britain. As Figure 4-2 shows, taking a sample of 824 surnames (see Winney *et al.* (2011) in Appendix 4 for the sampling process) and plotting the log of their 1881 frequency against the log of their MLQ demonstrates a strong relationship between the two. There is no single explanation as to why the

highest frequency surnames tend to have lower MLQ values. The conjecture here is that the plot in Figure 4-2 is the manifestation of three processes (approximately labelled A, B, C). In the case of very rare surnames (around A) the high MLQ values are unsurprising on the basis that they can physically only be present in a few areas (proximal or otherwise) and exist in family groups that will significantly increase their relative frequency. This appears to be the case with the surname "Soon" (Figure 4-3A) which has extremely high LQ values in only three dispersed locations. As the surname frequency distributions in Figure 3-10 demonstrate, this is likely to apply to the majority of surnames (but not the majority of the population).

Some caution needs to be exercised in this context because the LQ measure is more likely to produce extreme values when the number of surnames is small (Rogerson and Yamada 2009). Thus maps of rates can often give the perception of credible extreme values when they are nothing more than random variation. Mapping p-values, for example, can account for this to a certain extent but not entirely (Rogerson and Yamada 2009) so it should be left to the researcher to be aware of the limitations of small numbers in this context.

Most people in the population have a surname that falls around area B on the plot. These surnames should have a specific area of origin where the surname is most concentrated. It is from this area that the surnames are likely to have spread and become increasingly dispersed. This is illustrated by the surname "Metcalfe", in Figure 4-3, with the area of highest concentration surrounded by spatial units of increasingly reduced concentrations. Those with multiple concentrations, such as Wyer (Figure 4-1F), will also be included in this area of the graph, although they are not differentiable from the plot. Under these circumstances, measures of spatial autocorrelation (see next section) are useful.

The final aspect of Figure 4-2 relates to those surnames with the highest populations of bearers. In this section (labelled C) common Welsh surnames (Jones, Davies, Evans, Thomas, Hughes, James and Phillips), known for their relatively high populations and dispersion throughout Wales, are circled. In these patronymic cases, and also for occupational surnames such as Smith, the high levels of dispersion

(equated with low MLQ values) and frequency result from multiple origins. It follows that bearers of these surnames were able to multiply and spread more widely, thus aligning more with underlying population density than with any particular unique surname distribution.

The plot offers no insights as to why it may be the case that some surnames are more "successful" than others in terms of increased popularity or dispersion but it does confirm the clear population structure that this thesis seeks to unearth. Such processes have been statistically modelled by the likes of Manrubia and Zannette (2002) but no consistent explanations(s) for this process have been mooted.

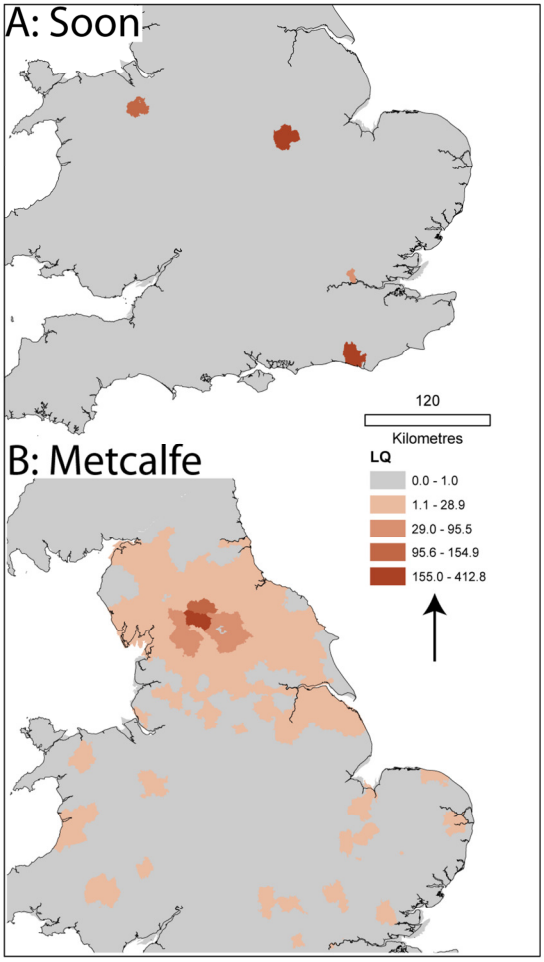


Figure 4-3: Example LQ distributions from two surnames identified in Figure 4-2 as being distinct from the main distribution of LQ vs. population values.

4.1.2.2 Measuring Spatial Autocorrelation: Moran's I

Spatial autocorrelation has been alluded to a number of times in this thesis and can be defined as the correlation of a variable (in this case surname frequency) with itself over space (Burt *et al.* 2009). It characterises the degree to which near objects are more similar to each other than distant objects (Tobler 1970). On this basis spatial autocorrelation is often positive, that is objects in similar locations have more similar attributes, but it can also be negative (objects appear more different the closer they are) or may not exist at all (Goodchild 1986). In this context the purpose is to establish the degree of spatial autocorrelation in a surname's distribution. The utility of this is twofold; firstly the areas of highest autocorrelation are likely to represent the core concentration of a surname (where spatial dependence is strongest). Secondly, it would supplement information provided by the LQ by hinting at the relationship between the areas of highest LQ and their neighbours. A number of measures have been developed to measure the degree of spatial autocorrelation in a dataset and here the widely used Moran's *I* statistic is deployed (Rogerson and Yamada 2009).

Moran's *I* can either be used globally, that is to establish the extent of spatial autocorrelation across an entire geographic area (in this case Great Britain), or locally to establish the degree of spatial autocorrelation within subsets of the data. Both measures have been tentatively utilized, with varying success, in the context of surname distributions (see Caravello and Tasso (1999)).

The global Moran's *I* statistic (Moran 1948) is one of the classic ways of measuring the degree of pattern in areal data (Rogerson 2006). It is calculated as:

$$I = \frac{m\Sigma_i^m\Sigma_j^m w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{(\Sigma_i^m\Sigma_j^m w_{ij})\Sigma_i^m(y_i - \bar{y})^2} \tag{4.2}$$

where there are m spatial units, y_i is the number of surname occurrences in spatial unit i, $\bar{y}$ is the mean of y_i (i=1, ,m), and w_{ij} is the measure of spatial proximity (as defined by a contiguity/weights matrix) between spatial units i and j. Positive (negative) values of I_i denote positive (negative) spatial autocorrelation (de Smith *et al.*

2009: 293), with an *a priori* expectation that values will rarely be negative and are unlikely to suggest very strong positive autocorrelation across all *m* given the regionally clustered nature of most surname occurrences.

It was thought that global Moran's *I* would provide a useful filter to remove those surnames that do not demonstrate a clear trend over space, and therefore cannot be classified as having a core area. However, investigations found global measures of spatial autocorrelation inappropriate in this context. Results from the measure suggested that surnames characterised by a highly clustered distribution within a few spatial units showed no spatial autocorrelation and therefore should be excluded. This is the case because the global nature of the measure tends to favour more gradual trends throughout Great Britain rather than a single area of high frequency. A very local surname will be restricted to a very small number (relative to the 220,000 OAs) of significant spatial units. It therefore makes sense to investigate local Moran's *I* in this context as it will identify the small areas over which a spatial relationship is apparent.

The local Moran's *I* statistic which is used to determine the existence of spatial autocorrelation around a specified subregion *i* (*i*=1, ,*m*) (Rogerson and Yamada 2009). Because most surnames form tightly clustered distributions focused upon a very small proportion of the spatial units in the study area this measure should not suffer the same limitations as its global counterpart. Using the same notation as Equation 4.2 it is defined as:

$$I_i = \frac{m(y_i - \bar{y})}{\sum_j (y_j - \bar{y})^2} \sum_i w_{ij}(y_i - \bar{y}) \tag{4.3}$$

Certainty in the result can be established by calculating the Z-scores associated with each value. The Z-score, or standard score, will indicate how many standard deviations the Moran's *I* value is away from the mean and therefore relates to the probability of its occurrence. Higher or lower Z-scores occur at the extremes of the frequency distribution and therefore have a low probability of random occurrence and a greater certainty associated with them.

Caravello and Tasso (1999) utilise local Moran's *I* to investigate different transitions in surname compositions that may reflect migratory process. By plotting the value of *I* against distance they produced correlograms on which they based a classification. From these, clinal transitions were identified along with processes such as isolation by distance (Caravello and Tasso 1999).

Figure 4-4 provides some example outcomes from preliminary investigations calculating local Moran's I according to Equation 4.3 and using the 1881 dataset. It is clear from these maps that the method has the potential to discern core areas of concentration in some cases. The range of Local Moran's *I* values is surprisingly small, suggesting very weak spatial autocorrelation, but the Z-scores associated with each value are extremely high. Based on these maps it is clear that Moran's *I* could provide a good visual indicator in this context, but that it would be extremely hard to distinguish between what is significant and what is not using such a small range of values. In addition, if a spatial unit is assigned a high value of *I* then others surround it with similar frequencies of that surname. There is no indication of whether the number of surname occurrences within these areas is high or low, so areas with few occurrences may be assigned the same values of *I* as areas with many occurrences. Further investigations suggested that the local Moran's *I* statistic was suitable for detecting some surname localities, but that it did not identify the clusters associated with rare surnames or regionalise many common names. For the purpose of defining a surname's core area, only those areas with high surname frequencies are of interest; thus in order to differentiate between high and low occurrence areas, it was thought possible to use a combination of local Moran's *I* and the LQ.

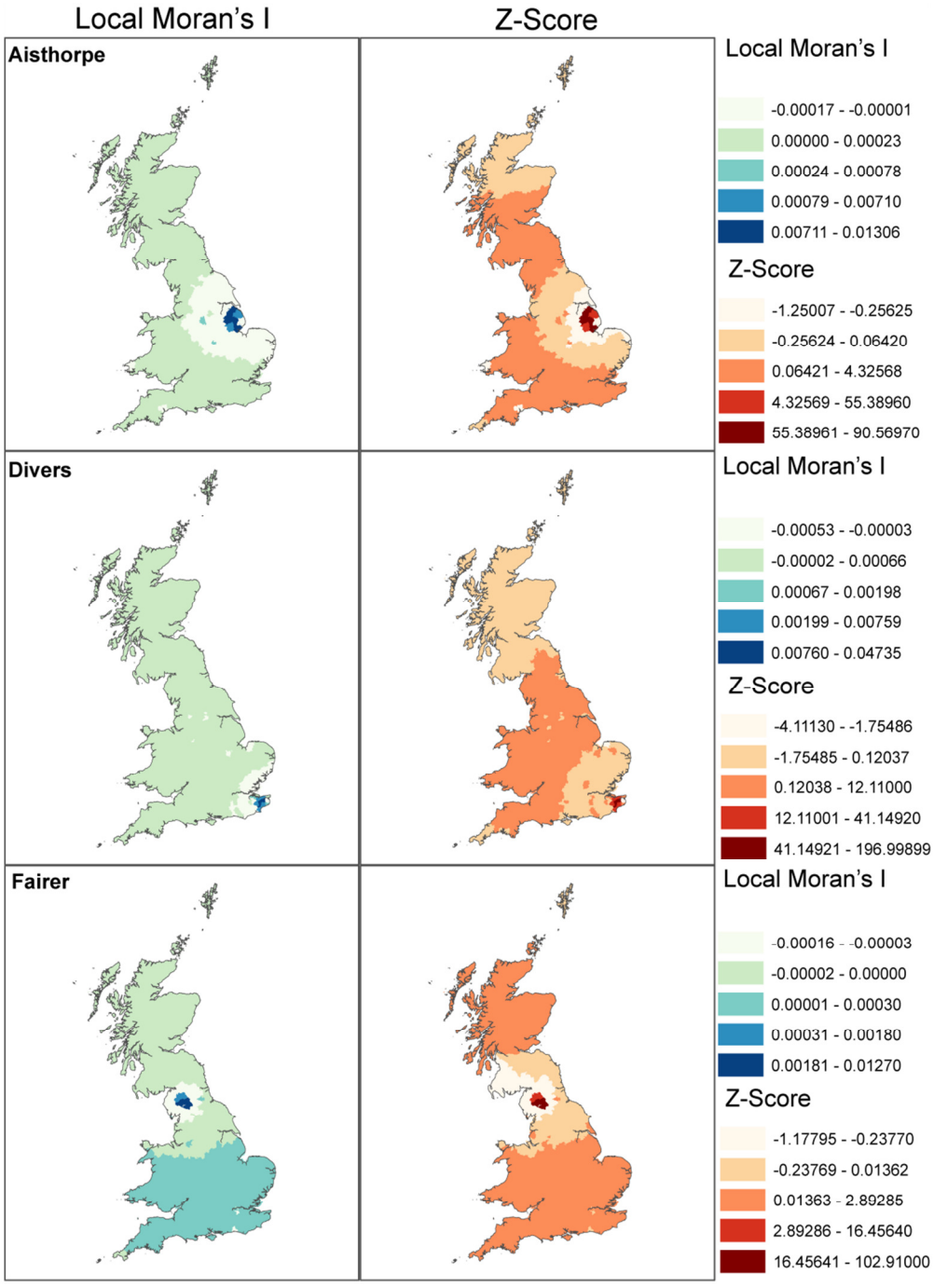


Figure 4-4: Three example outcomes from the preliminary calculations of the local Moran's *I* statistic. Darker blues represent areas of higher spatial autocorrelation and signal the surnames' core area of concentration. Darker reds signal higher certainty in the result based on the Z-scores.

4.1.3 APPROACHES USING POINT DATA

4.1.3.1 Kernel Density Estimation

The concept of modelling the density of population characteristics is well established. Kernel based approaches have been successfully used for a range of applications, most notably crime mapping (Chainey *et al.* 2008). Amongst the most well-known applications beyond crime are the national population models of England, Wales and Scotland (Bracken and Martin 1995), which have been extended to automatic neighbourhood identification procedures by Martin (1998a).

Kernel density estimation (KDE) has become a standard technique for characterising spatial variations in point densities (Lloyd 2007). The purpose of KDE is to estimate the intensity of a point process at a set of given locations on a gridded surface (Rogerson 2006). It has a variety of applications, including smoothing, interpolation of continuous surfaces from point data, probability distribution estimation and hotspot detection (de Smith *et al.* 2009). The intensity of the process is defined as the number of points per unit area as the area around the given location decreases in size (Rogerson 2006). Here, KDE is used to estimate the density of occurrences of a phenomenon, in this case surnames, across Great Britain. KDE proceeds by placing a kernel over each node of a regular grid. Each cell on the grid is assigned a density estimate that is the sum of the kernel values within its locality (as defined by the bandwidth) divided by the total area of the locality from which the values are drawn.

Observed occurrences are assigned a weight according to the kernel function $k(d_{ij})$ which is a function of the distance from the grid i point to the observation location j and takes the form:

$$\hat{\lambda}_i = \sum_{j=1}^{n} k(d_{ij}) \qquad (4.4).$$

The extent of a kernel's influence is determined by two things: type and bandwidth (h). The most commonly used kernel type has an (unbounded) Gaussian (normal)

distribution (Kelsall and Diggle 1995). The primary effect of the use of an unbounded kernel is the production of a slightly more generalised surface because there is a less abrupt reduction in the influence of each occurrence on the surrounding grid nodes. A one-dimensional representation of this process can be seen in Figure 4-5.

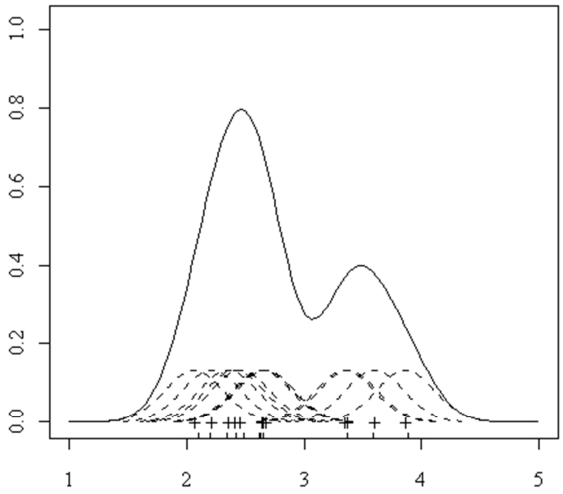


Figure 4-5: A one dimensional representation of KDE. Crosses are occurrences and dashed lines represent normal kernels placed over them. The solid line is the final estimate and is the sum of the underlying dashed lines. The value assigned to each grid cell is taken from the point on the solid line directly above the centre of the grid cell. (Source: R Development Core Team 2011).

Bandwidth (*h*) typically has a more marked effect upon the density estimation than kernel type and is defined as the extent of the area around each grid cell from which the occurrences and their respective kernels are drawn. Larger bandwidths will encompass more points and therefore produce more generalised estimates than those using a smaller bandwidth. It is also possible to vary bandwidth in order to accommodate local variations in each point distribution. This can be determined by user-specified selection criteria such as the minimum number of occurrences to include within a circle centred over each occurrence (de Smith *et al.* 2009). Fixed bandwidths (that do not change during the KDE calculation) are generally used to represent high relative incidence within a global distribution.

Whilst the bandwidth's importance can be over-stated (Scott 1992), its impact on the resulting density estimation is clear. A small bandwidth heavily weights information in close proximity to a grid cell creating an estimate with little bias but a high degree of uncertainty (Rogerson and Yamada 2009). Conversely a larger bandwidth will produce more certain estimates but suffer from an increase in bias because more nonlocal information is being used (Rogerson and Yamada 2009). Bandwidth selection is often subject to a degree of trial and error (de Smith *et al.* 2009). Bowman and Azzalini (1997: 31-35) list three of the most effective bandwidth types to be optimal smoothing, normal optimal smoothing and cross-validation. After initial testing, normal optimal smoothing was chosen to be the most appropriate in this context. This was because it offered a more conservative approach than straightforward optimal smoothing and was much quicker to calculate than cross validation (an important consideration given the size of the datasets used here). Normal optimal smoothing is based on the premise that smoothing parameters (bandwidth) should decrease with sample size (n) proportionately to $n^{-1/6}$ and has the advantage of producing a smoothing parameter that requires very little calculation (Bowman and Azzalini 1997). The normal kernel is one of the smoothest distributions so the optimal bandwidth value will be large and presents the risk of over-smoothing non-normally distributed data (Bowman and Azzalini 1997). This can be avoided through careful parameterisation that, as Section 4.1.5 below shows, reduces the bandwidth estimate to under-smooth the surface.

As previously mentioned, it is a feature of KDE that density estimates are provided over the entire grid, and not just confined to the specific areas where surnames are found. In addition, there is no need to define a priori thresholds, as is the case with the discrete methods outlined above. Users can simply select their desired density thresholds afterwards. This produces a slightly fuzzier impression of the distribution of surnames than would be the case with area-based methods. In the context of inferring a surname's origin from its core areas this is considered advantageous because it goes some way to account for the fact that people are likely to have made local migrations within their present areas.

Some example KDE surfaces are provided in Figure 4-6 and produced from the 1881 Census data. All four show the effectiveness of KDE for capturing the spatial distributions of surnames. The density surfaces can be visualised in two or three dimensions and can provide the inputs to standard surface analysis techniques such as establishing gradients or drawing contour lines.

KDE requires careful parameterisation in order to produce results that are closely representative of a surname's spatial distribution. It is important, for example, to select a kernel function that derives a surface that encloses a volume equal to the total number of occurrences in the spatial distribution. This prevents the density

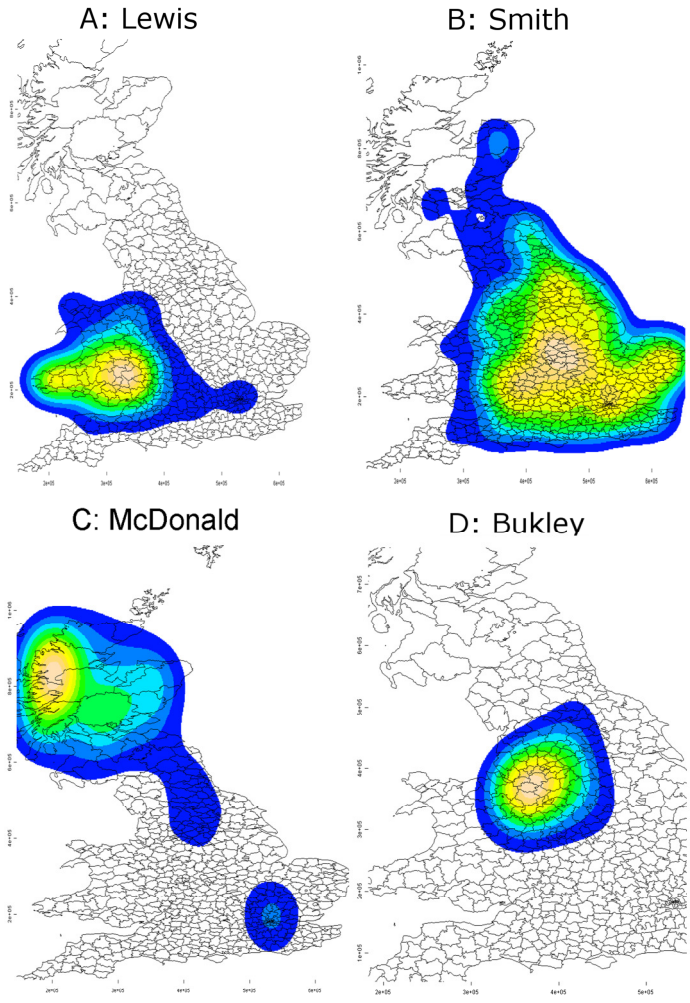


Figure 4-6: Example KDE surfaces produced using surname frequencies obtained from the 1881 Census. Even for common surnames, such as Smith, a clear spatial pattern is present and captured by the KDE.

surface from implying there are more occurrences than reality (Lloyd 2007). As alluded to earlier, KDE relies on a number of subjective decisions about the best parameters to select. It is therefore important to undergo a thorough process of validation before the final density estimate is produced.

4.1.3.2 Discontinuities

In addition to KDE there are a number of other surface based approaches that may be appropriate. These vary from relatively simple regression surfaces or inverse distance weighting (IDW) through to complex geostatistical procedures such as Kriging (see Chapter 6 of de Smith *et al.* 2009). The focus here is not upon the interpolation methods themselves but the appropriateness of a surface based on surname counts (rather than density) for the identification of areas of high concentration. Once a surname population surface has been created there are a number of ways to automatically identify transitions between the peaks associated with high frequencies and the troughs associated with low frequencies.

One of the simplest methods to identify areas of the population surface that represent particularly high or low frequencies is to partition it using contour lines drawn at specific frequency intervals. This approach has been used before by Sokal *et al.* (1992) and can provide an effective overview of the surname's distribution. The method may be improved by looking for discontinuities in the surface that are likely to result from more abrupt transitions from high-frequency to low-frequency areas. These will represent natural "breaks" in the surname's distribution and provide an alternative to drawing contours on a surface at arbitrary intervals. Surface wombling, or boundary analysis, is the formalisation of this process and has been used in a number of contexts including in surname frequency analysis (see Sokal *et al.* (1992)). The purpose is to identify the points of steepest ascent or descent on a fitted spatial surface as these represent the areas of the most abrupt variations (Lu and Carlin 2005). With many surface-based techniques there is a risk of over-smoothing the more abrupt transitions in surname frequency. This is important because such transitions are likely to mark areas in the distribution of most interest (Manni *et al.* 2004). For example, boundaries often occur in and around urban areas as particular

migrant groups cluster together, or along national and linguistic borders (see Chapter 6 for more evidence of this).

Brunsdon (2009) calls these abrupt transitions in population data "Social Faultlines" and has developed a method of exaggerating them for identification purposes whilst maintaining the general trends in the data. The goal is to provide a smoothing approach that smoothes when needed, but does not smooth over discontinuities. The first aspect of the method requires the creation of a gridded population surface. Here, as in Brunsdon's (2009) example, kernel regression is used but any other interpolation technique, such as IDW, would also be appropriate. Kernel regression estimates a trend surface using a set of point locations (*x,y*) based on their attributes(z) by placing a kernel around each point (*x,y)* and taking a weighed mean of z with the weight decreasing with distance from (*x,y*). The function used here is defined as:

$$w = \exp\left(-\frac{d^2}{2k^2}\right) \tag{4.5}$$

where *d* is the distance, *w* is the weight assigned to each z value and k is the smoothing parameter (larger values produce a smoother surface).

The second stage is concerned with exaggerating the discontinuities using an adapted version of anisotropic smoothing (Brunsdon 2009). Anisotropic smoothing was developed in image processing in order to detect "edges" between pixel values that represented objects within the image (see Perona and Malik (1990)). The approach replaces the cells of a grid with the weighted average of their neighbours as follows:

$$z *_{ij} = \frac{w_{1,0}z_{i+1,j} + w_{-1,0}z_{i-1,j} + w_{0,-1}z_{i,j-1} + w_{0,1}z_{i,j+1}}{w_{1,0} + w_{-1,0} + w_{0,-1} + w_{0,1}} \tag{4.6}$$

where the asterisk denotes an updated value (Brunsdon 2009). To increase the smoothing it can be repeated multiple times. In its basic implementation the weighting does not vary (indices *i* and *j* are stationary) and therefore can over-smooth in the same way as surface-based techniques. Brunsdon's (2009) adaptive version,

however, adjusts the weighting according to the slope (difference in values between adjacent cells) in the following way:

$$z *_{i,} = \frac{w_{i+1,j} z_{i+1,j} + w_{i-1,j} z_{i-1,j} + w_{i,j-1} z_{i,j-1} + w_{i,j+1} z_{i,j+1}}{w_{i+1,j} + w_{i-1,j} + w_{i,j-1} + w_{i,j+1}} \tag{4.7}$$

where

$$w_{i,j} = f(s_{i,j}) \tag{4.8}$$

and $s_{i,j}$ is a slope estimate at pixel *(i,j)*. *f* is a decreasing function, in this case:

$$exp\left(-\frac{1}{2}\frac{s_{i,j}^2}{h_2^2}\right) \tag{4.9}.$$

Steeper slopes are assigned lower weights and vice versa, thus creating the desired effect of preserving the most abrupt transitions in the surface. Slope in this case is based on the assumption of equal grid cell spacing and is defined as:

$$s_{i.j} = \frac{(z_{i+1,j} - z_{i-1,j})^2 + (z_{i,j+1} - z_{i,j-1})^2}{4} \tag{4.10}.$$

Some smoothing is still applied to the rest of the surface so it is acceptable to under-smooth the input surface in order to account for this (Brunsdon 2009). It is then possible to draw contour lines along the base of the discontinuities to visualise and partition the surface.

Figure 4-7, produced with the 1881 Census, offers some examples of the results from the social faultlines approach in this context. Kernel regression was used to produce the 150 by 150 grid of surname frequencies and then "smoothed" with 50 iterations of the adapted anisotropic diffusion. The preliminary results have been visualised as pseudo-3D perspective plots and with contour lines on a base map. The approach appears effective for less common surnames and correctly identified their areas of highest frequency. It was, however, less effective with more common surnames, "Smith" in Figure 4-7 being the extreme example. This reflects the impact of the underlying configuration of spatial units. The reason for a peak occurring over the area of Glasgow is simply that it is a Registration District with a very large population

and will likely have more Smiths than surrounding less populated areas. In less populated areas there may be more Smiths but they have been partitioned into a greater number of spatial units thus reducing their average frequency. The approach also failed to account for the impact of larger numbers of points associated with the greater numbers of spatial units in more populated areas. In these cases there will be greater certainty in the interpolation but the combination of density and frequency will have less of an impact than with other techniques, such as KDE.

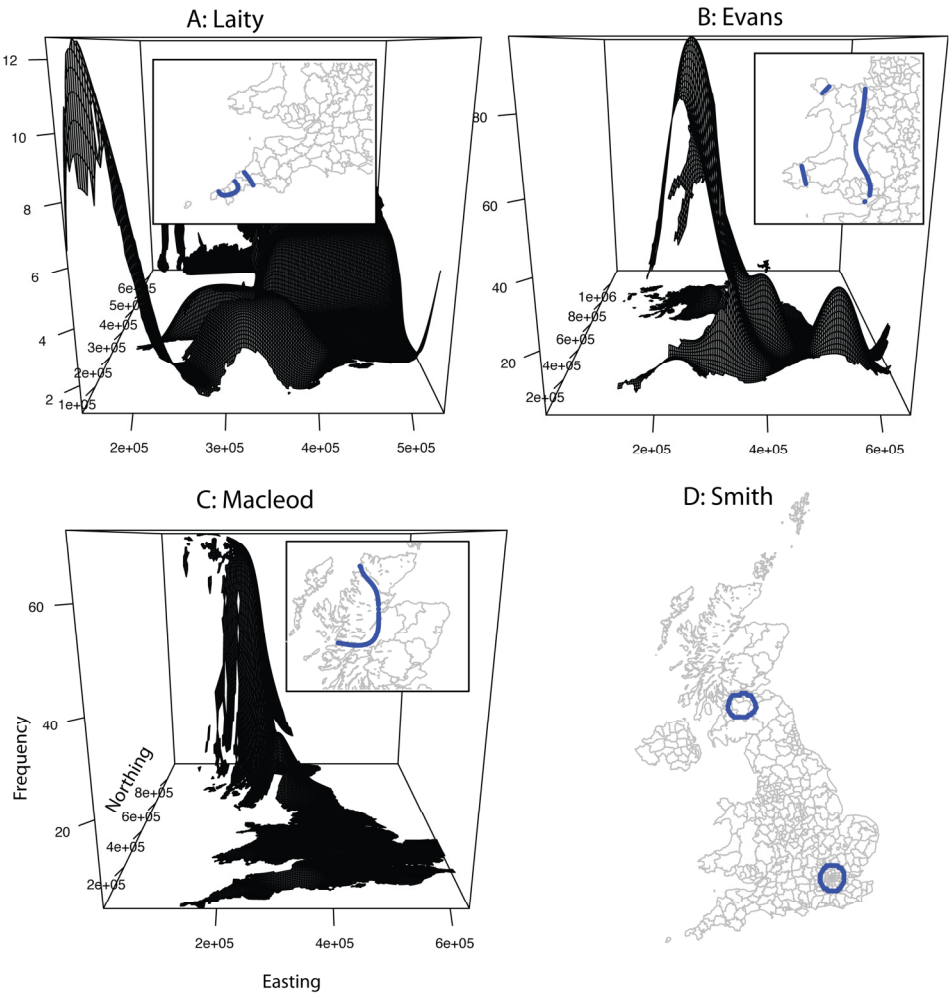


Figure 4-7: Results from the discontinuity preserving anisotropic smoothing shown in pseudo-3D (with the exception of Smith due to the complexity of the surface) and conventional map form with a contour line. The discontinuities occur where there is a sudden reduction in surname frequency and may indicate the outer limits of a surname's core concentration. With the exception of Smith these results correctly identify the areas of highest concentration for each surname. Best viewed digitally: http://spatial.ly/igvcRY

4.1.4 REQUIREMENTS FOR A TYPOLOGY AND SELECTED METHODOLOGY

The four methods outlined above vary in their complexity and each capture slightly different aspects of a surname's spatial distribution; their utility is therefore applications dependent. In this case the application is the creation of a surname typology for Great Britain that can be used as a basis for future research. For such a typology to be widely adopted it requires a number of features. Firstly, it needs to apply to the complete distributions of a large number of surnames in Great Britain – not simply partial distributions selected for convenience in particular surname instances, as has been the case to date. The effects of the underlying distributions of population also need to be accounted for, in order to ensure that the core concentrations identified reflect more than population density.

Secondly, and most importantly in the context of population genetics, a typology needs to distinguish between surnames that are monophyletic (single area of origin), polyphyletic (multiple areas of origin), or have no specific origin. As was outlined in Chapter 2 this provides an important method of assessing the genetic and cultural relatedness between individuals with the same surname. In addition, the typology should be flexible in terms of its identification of the contemporary areas of highest concentration. In order to sample specific groups within a population, it makes sense to target those areas where they are most concentrated. If, however, a sufficiently large sample cannot be obtained from a specific target area it can be expanded. In the past this has been based on arbitrary distance from a specific point (see Kaplan and Lasker (1983)). The typology created here represents an improvement by following surname frequency gradients rather than arbitrary buffer operations. This requirement alone suggests a surface-based approach may be more appropriate.

A further consideration, and one that also favours the use of surface-based methods, is the requirement for temporal comparisons. Historical datasets rarely have consistent spatial units and therefore restricts direct comparisons using the original spatial units. The advantage of surface based approaches, such as KDE, is that they are distribution free (Martin 1996): that is, the size and configuration of the input

areal units is relatively unimportant. For these reasons, the impact of inherent clusters produced by the configuration of administrative units and their effect on computational efficiency is greatly reduced. With the appropriate grid resolution, however, much of the information contained within densely populated areas need not be lost and the resulting increase in data redundancy in more rural areas is not considered problematic. The removal of inherent spatial clusters (that persist for both absolute and relative numbers of occurrences) is important, especially in the context of historical inferences about populations that were less urbanised than today. The additional advantage is the decreased dependency of the results on consistent spatial units when using historical data. For example UK Census Output Areas (OAs) have only existed since the 2001 Census and they do not tessellate well with the parishes of the 1881 Census. The adaptability afforded by converting the data into a gridded representation therefore provides a pragmatic solution to inconsistent spatial units. In addition, the nature of the data within the typology should also be readily integrated with and compared to spatial and non-spatial datasets such as electronic gazetteers.

The final consideration is more pragmatic and relates to the limits placed on the computation of the metrics in the typology. The metrics need to be calculated for tens of thousands of surnames that, in the case of the 2001 data, have been geocoded to fine-scale spatial units. It would not be practical to implement an approach that took more than a couple of seconds to run for each surname.

In the light of the above requirements KDE provides the most promising method for typology creation. Use of threshold LQ and Moran's *I* values transpired to be inappropriate because some surnames occur in very small areas whereas others tend to exhibit much less concentrated distributions. In the former case it is probable that there will be few surrounding spatial units with similar values, resulting in a low value of *I* but a high LQ. In the latter case there may be many spatial units grouped together with similar values that result in high values of *I* but lower LQ values. This problem is exacerbated by surnames with multiple core areas of concentration because each may have different characteristics. In addition the calculation of local measures of spatial autocorrelation is excessively computationally intensive, primarily

because of the size of the contiguity or weights matrix required. If, for example, 2001 OAs were used, a matrix of approximately 220,000 by 220,000 would be required and re-weighted for every surname, thus exceeding practical computational limits. In addition, the calculations are sensitive to the definition of contiguity used (Bivand *et al.* 2008).

A further limitation to the use of discrete areal units, and one that also partially affected the social faultlines approach, is widely known and concerns the configurations and spacing of spatial units within Great Britain. The use of spatial units representing approximately standardised population sizes, such as OAs, is advantageous because it minimises the impact of small numbers in sparsely populated areas whilst reducing the disproportionate generalisation of densely populated areas. As Figure 4-8 illustrates, to produce spatial units of consistent population size, their geographical extent varies such that urban areas with dense populations have much smaller OAs than rural areas with sparse populations. Any surname data attached to these spatial units is therefore inherently clustered around urban areas. In addition, because this research is premised on the assertion that surnames (and by implication, their bearers) tend to remain in ancestral areas, it would be unreasonable to assume that such heartlands exist within the arbitrary administrative boundaries to which population geography research is often constrained. A surface based approach is less sensitive to these issues.

Compared with the other methods, especially Moran's *I*, KDE is quick to compute and not bound by specific administrative geographies. If the surface needs to be partitioned to identify the area of highest concentration according to a threshold value, this can be achieved through use of a contour line. The threshold can be increased or decreased to match the surname sampling requirements of the particular study. This is not possible with the discontinuity surfaces as they only reflect abrupt boundaries and not the gradual transitions common to many surnames.

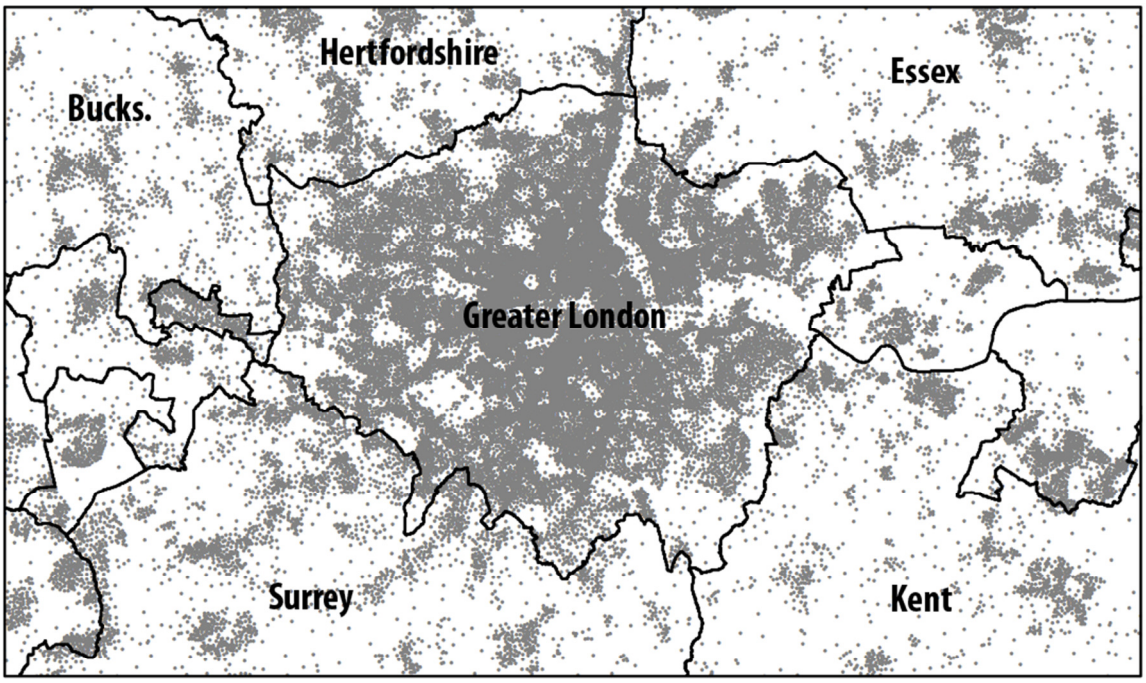


Figure 4-8: A demonstration of the uneven distribution of OAs around Greater London. Grey points represent OA centroids.

4.1.5 FINAL METHODOLOGY

The above section sought to demonstrate a number of the methods trialled to create an automated surname typology for Great Britain. Its purpose was to provide some insights into the choice of KDE as the selected method and some of the potential issues to overcome if the methodology is to be successful. What follows are the detailed methodological steps and processes of parameterisation behind the final approach. The data were georeferenced to population-weighted centroids of OAs in the case of 2001 and geometric centroids of Registration Districts in the case of 1881.

The methodological steps taken to produce the metrics in Table 4-1 are summarised in Figure 4-9 and articulated below. They were devised following a thorough investigation (based on a 10% sample of surnames with a range of frequencies) into

Table 4-1: A full list of metrics provided by the KDE classification. Published in Cheshire and Longley (2011a).

Variable Name	Description
Surname	Surname
Unique_ID	Unique ID assigned to each surname
Surname_Frequency	The number of bearers of that surname in Great Britain.
Number_of_Cores	The number of surname cores.
Core_Number	The core number. Surnames with multiple cores will be given multiple rows: 1 for each core.
X_Coordinate	The X coordinate for the centre of the core. This uses the British National Grid.
Y_Coordinate	The Y coordinate for the centre of the core. This uses the British National Grid.
Core_Area	The area (in km²) of each core.
Toponym_ID	The unique ID (from the GB Gazetteer) of the toponym. Defaults to "NULL" if there is no feature in the gazetteer that matches the surname. "Not close to placename" means a feature match exists, but it is outside the surname core.
Toponym_Name	The place name of the topopnym. Defaults to "NULL" if there is no feature in the gazetteer that matches the surname. "Not close to placename" means a feature match exists, but it is outside the surname core.
Toponym_X	Toponym X coordinate (British National Grid). Defaults to "NULL" if there is no feature in the gazetteer that matches the surname. "Not close to placename" means a feature match exists, but it is outside the surname core.
Toponym_Y	Toponym Y coordinate (British National Grid). Defaults to "NULL" if there is no feature in the gazetteer that matches the surname. "Not close to placename" means a feature match exists, but it is outside the surname core.
Toponym_Feature_Type	The feature type code.
Core_Distance_1	Distance between 1st and 2nd cores. Defaults to 0.1 if there is 1 core.
Core_Distance_2	Distance between 1st and 3rd cores. Defaults to 0.1 if there are fewer than 3
Core_Distance_3	Distance between 2nd and 3rd cores. Defaults to 0.1 if there are fewer than 3
Core_Distance_4	Distance between 1st and 4th cores. Defaults to 0.1 if there are fewer than 4
Core_Distance_5	Distance between 4th and 3rd cores. Defaults to 0.1 if there are fewer than 4
Core_Distance_6	Distance between 2nd and 4th cores. Defaults to 0.1 if there are fewer than 4
Surname_Pop_Within_All_Cores	Total number of surname bearers within all cores for that surname.
Surname_Perc_Within_All_Cores	Percentage of the surname's bearers within all cores for that surname.
Surname_Pop_Within_ Core	Number of surname bearers within specific core.
Surname_Perc_Within_Core	Percentage of surname bearers within specific core.
Relative_Concentration	The contour value required to include the desired Surname_Perc_Within_Core.

the most suitable combination of parameters to ensure the effective handling of the breadth of spatial characteristics exhibited by surnames. Such is the range of surname frequencies that the effect of each parameter may differ for frequent surnames (which are likely to be more dispersed) compared to rare surnames. The data were therefore split into quantiles based on the surname frequency distribution. 10% of the surnames that occurred within each quantile were randomly sampled so that each quantile was represented by approximately the same number of surnames. Table 4-2 shows the number of surnames (sample size) from each quantile and their frequency of occurrence.

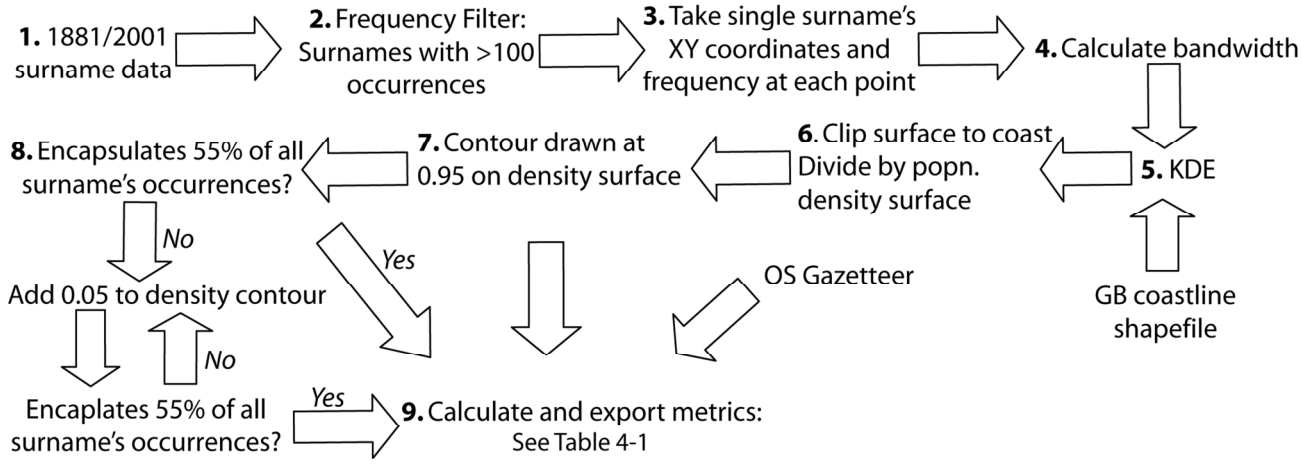


Figure 4-9: Flow chart to illustrate the methodological steps taken to produce the metrics outlined in Table 4-1. Published in Cheshire and Longley (2011a).

As outlined above, bandwidth selection (step 4, Figure 4-9) has an important impact on the results. Here the bandwidth is specified in the *x* and *y* dimensions using normal optimal smoothing for each (Bowman and Azzalini 1997). Altering the bandwidth in two dimensions proved a pragmatic solution to the physical constraints placed by the coastline on Britain's population distribution. In this context a fixed bandwidth was considered more appropriate because the use of variable bandwidths created additional hot-spots in towns and cities: such concentrations of population were largely absent when the names were first coined in history, and the objective was to represent high relative frequencies over more geographically extensive areas.

Table 4-2: Details of the validation sample taken from the full dataset. Sample size represents the number of unique surnames and mean frequency is their number of occurrences.

Quartile	Sample Size	Mean Frequency
1	678	130.6
2	695	226.2
3	687	516.2
4	506	24188.9

It would not be practical to report the bandwidths used for each surname on an individual basis: instead, Figure 4-10 shows the relationship between the mean bandwidths (from the x and y dimensions) and the surname's frequency in the 9 Government Office Regions of Britain (GOR), Wales and Scotland. This distribution

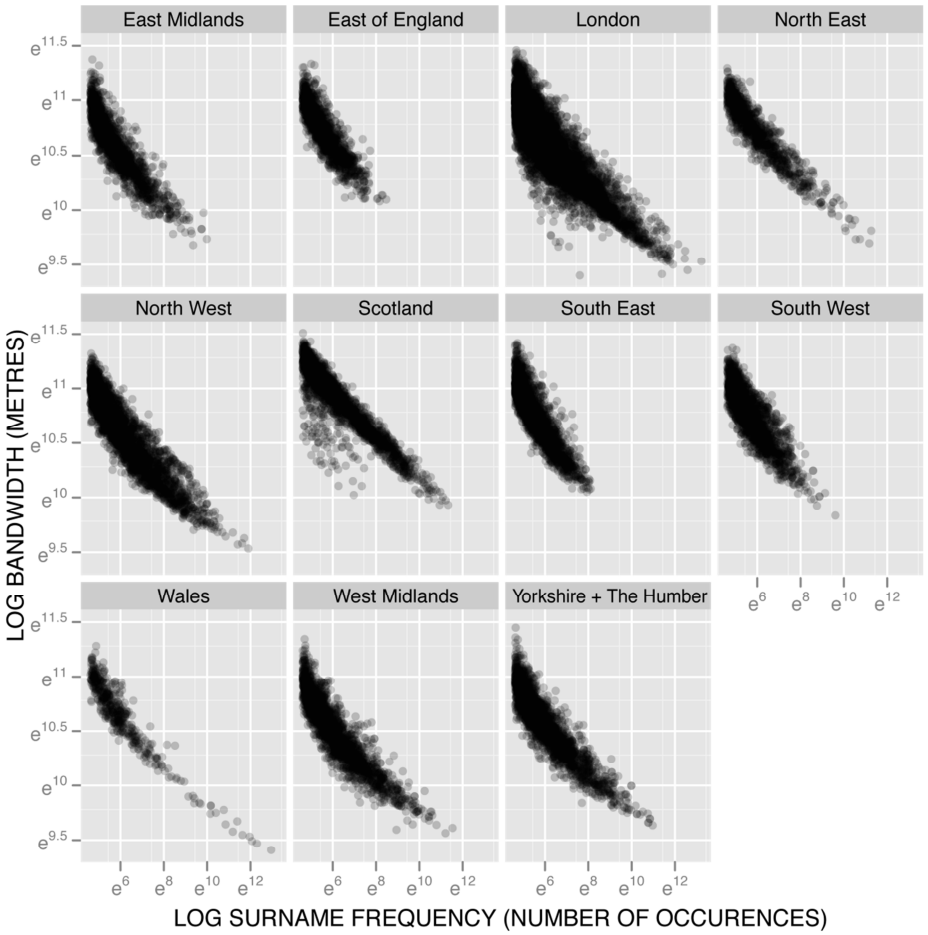


Figure 4-10: A plot showing the relationship between the mean bandwidth (h) (calculated with normal optimal smoothing) and the frequency of surname occurrences. Published in Cheshire and Longley (2011).

is unsurprising because, as was outlined above, the use of normal optimal smoothing reduces bandwidth in proportion to sample size (in this case number of OAs containing the surname). It does, however, provide an indication of the range of bandwidths required to adequately capture the individual distributions of surnames in Great Britain.

Finally, the frequency of surname occurrences also informs the choice of bandwidth. Here rare surnames (those with fewer than 100 occurrences) have been excluded for a number of reasons. Rare surnames require much larger bandwidths relative to the density of their distributions, resulting in a greater relative spread in the KDE in comparison with many moderate frequency surnames that attain higher densities and thus can be represented using a tightly defined KDE (Bowman and Azzalini 1997). In addition, a point is reached where there is little to gain in undertaking a KDE. The majority of rare surnames cluster in specific areas of Great Britain and automatic detection of these can be achieved using more straightforward methods such as mapping the location quotients, as outlined above. These reasons, combined with the relative importance of chance scattering of rare name bearers, contributed to the decision to only include surnames with over 100 occurrences in our analysis.

For the KDE surfaces to meaningfully represent a surname's core concentration in Great Britain, variations in population density play an important role. This is most evident with respect to urban areas where there is an increased probability of any surname occurring. For example, the occurrence of 10 Cheshires in an OA with a population of 250 is half as important in representing the distribution of this name (alone) than the same number in an OA with a population of 500. One possible solution is to weight each OA centroid by the relative frequency of that surname (OA surname frequency/ OA total population). This proved ineffective, however, because the high concentration of small spatial units in urban areas was sufficient to counter any weighting provided by the relative frequency measure. A more successful approach calculated two density estimates – one weighted by the total population in each OA and the other weighted by a surname's frequency in each OA. By dividing the surname KDE by the total population KDE (step 6 of Figure 4-9) it is possible to reduce the spatial structure effect created by densely packed urban areas. A

balance needs to be struck when accounting for densely populated areas; over-compensation will lead to legitimately "urban" surnames (such as toponyms, or migrant surnames) being redistributed towards less well-populated areas where they nevertheless occur in relatively low numbers. Conversely, under-compensation will over-assign surname cores to urban areas, something that would lead to the majority of surname cores being centred on Greater London.

As mentioned above, the 2001 Electoral Register was used in the parameterisation process. In this case the surname frequencies were spatially assigned to population-weighted centroids of each of the 218,038 OAs in Great Britain. The large volume of points required a balance between processing time, which increases with the grid resolution, and the level of generalisation that has an inverse relationship with resolution. After experimentation, 16,900 4 km by 4 km cells on a 130 by 130 grid provided an acceptable compromise between the computation time for each KDE and the level of detail generated. Geographic (rather than grid) distances are used in the KDE calculation. Preliminary testing showed that reducing the grid size to less than 100 by 100 began to have a detrimental impact as the surfaces, and the associated contour lines drawn along them (see below), became over- generalised. After the KDE was calculated the values were standardised (of between 0 and 1) as follows

$$z = \frac{a - a_{min}}{a_{max} - a_{min}} \tag{4.11}$$

where *a* is a matrix of all the values on the grid. This enabled direct comparisons to be made between different surfaces. A density value of 0.95 and above therefore represents the top 5% of the density distribution and likewise a density value of 0.10 and below represents the lowest 10% of the density distribution. The final step of the KDE calculation clipped the grid to the British coastline. This also provided a pragmatic response to edge effects and ensured the surname distributions were plausible. Had a volume preserving KDE been implemented this may have been problematic, but in this case the surface is simply an indication of relative density. Population counts are provided by actual data points (not inferred from the surface) and these, of course, all occur on the landward side of the coastline.

Examples of the resulting density surfaces are provided in Figure 4-11. The surnames included represent the diversity of patterns produced by the KDE ranging from tightly clustered names such as Bamber, more dispersed surnames such as Palin, surnames with multiple clusters such as Khalil and those with secondary clusters in urban areas such as Macleod.

KDE produces a density surface onto which a contour line can be drawn at a specified threshold (step 7 Figure 4-9) and, by straightforward extension, the density surface can be used to identify a contour that encloses a pre-specified percentage, of

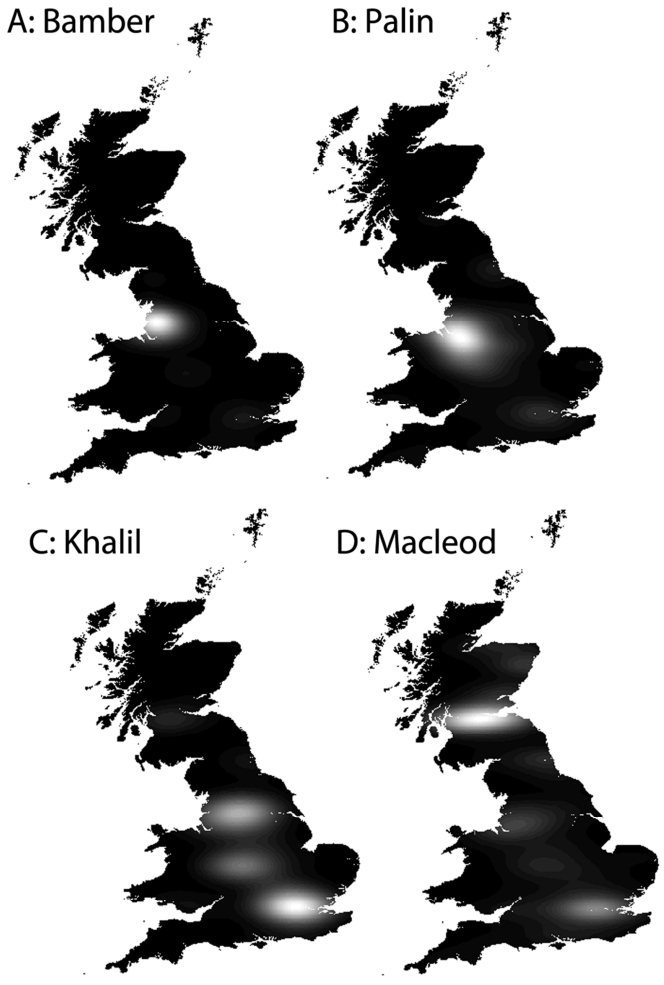


Figure 4-11: Illustrative KDE surfaces for Bamber, Palin, Khalil and Macleod. Produced from the 2001 Electoral Register data. Published in Cheshire and Longley (2011).

the surname's total population. Drawing a contour line around the areas of highest density at a pre-defined threshold, for example around the top 30% of the density range, will identify the area where the surname is most concentrated. This area, however, may only include a few of the total number of surname occurrences. An alternative approach uses the density surface as a guide and expands the contour (by decreasing the density threshold value) until a pre-specified percentage of the surname's total population is contained within the contour. Figure 4-12 illustrates the different outcomes from each approach. It clearly shows that the spatial extent of each surname core is dependent on the whether a core is defined according to concentration or population. After much consideration both approaches were used to produce the final set of metrics to describe a surname's spatial distribution.

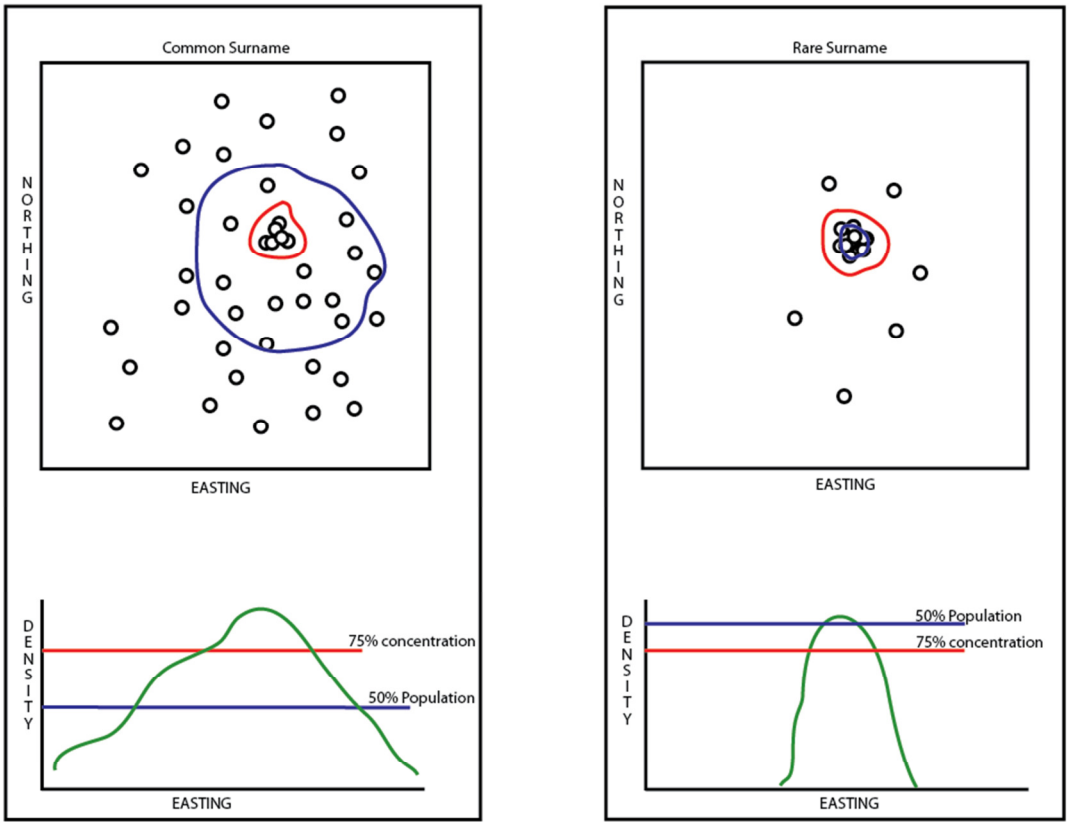


Figure 4-12: A diagram to show the different core outcomes obtained from a population threshold value when compared with a density threshold value. The point distributions represent how the surname occurrences appear on the grid for a common surname (left) and a rare surname (right). The lower plots show the likely KDEs for each. In the case of common surnames the population threshold produces a larger core area. The opposite is true for the rare surname example.

When seeking to infer a surname's origin from its area of concentration, however, identifying the area of highest density was considered most useful because in many cases the contemporary distributions of surnames had spread beyond their core areas to towns and cities, for example, causing the core region to span beyond the historical "hotspots". For this reason it was used for two additional parameters: "Area of Origin" and "Number of Areas of Origin". The population based threshold and density threshold surfaces were parameterised slightly differently as outlined in more detail below. The surface used for the density-based contour was produced using a bandwidth calculated according to the criteria suggested by Bowman and Azzalini (1997). A threshold value of 0.95 was used because the purpose was simply to identify the area of highest concentration. Lower values would produce larger contours that are also likely to be more numerous.

The contour line drawn around 55% of the surnames' population was used for the majority of metrics as it makes most sense to deal in actual proportions of people as opposed to kernel density values. Setting parameters for the population-based contour was a little more complex. It still relies on a density surface and density threshold values but the latter are iteratively decreased until a pre-specified proportion of the surname's population is contained within the core. Initially, the population-based threshold is drawn at a value of 0.95 (the highest 5% of the density surface) and the proportion of a surname's occurrences within it are calculated. If the proportion exceeds the threshold value, of for example 55%, the core is unchanged for the rest of the analysis. If, however, less than 55% of the surname's occurrences are contained within the core then the contour is drawn at the lower value of 0.90. The process iteratively reduces the contour value by 0.05 until it reaches 0.10. If 55% (or some other desired value) of the surname's population is still not enclosed at this value the surname is considered too dispersed for having a core area. Concentrated surnames will have cores that are drawn around higher density values and less concentrated names will require lower density values. This incremental growth along the density surface permitted the use of a smaller bandwidth when producing the surface. This is possible because a density threshold would highlight each of the peaks in the distribution, thus increasing the likelihood of highlighting multiple cores that are in fact smaller fluctuations within a larger core. Such fluctuations are unlikely

to affect the final contour assignment because the density threshold value is likely to decrease to the point where they will be highlighted as a single core. An additional advantage to this approach is the preservation of abrupt discontinuities. As discussed above, larger bandwidths smooth the data and may generalise abrupt transitions.

The dimensions of a surname's core population, as characterised by the population-based contour, will therefore be heavily dependent on the selected threshold value. In this case the value was informed by 10 validation runs that incrementally reduced the population threshold by 5% from 95% to 40%. The effect of the threshold percentage on the total core area of the surnames can be seen in Figure 4-13. Aside from the expected increase in outlier surnames classified as having very large core areas in quartiles 1 and 4, an interesting pattern emerges in these plots. Within each quartile sample the first (lowest) quartiles of each area remain consistent with each increase in the percentage of the population included within the core. The interquartile range therefore increases as a result of a larger variation in the size of core areas that are greater than the median. This generally occurs at threshold values of greater than 55% and is likely to reflect the increase in surnames classified as having more than one core, as shown in Figure 4-13. For this reason, 55% was considered a reasonable population threshold value to use.

The availability of electronic gazetteers facilitates partial validation of the areas generated for toponymic surnames. After establishing a surname's core locality the surname was used to search the Ordnance Survey (OS) Gazetteer (Ordnance Survey 2010). If there is no match then the name is not considered toponymic. Where single or multiple matches are found between the surname's core area and the gazetteer, locational information from the latter is used to perform a point in polygon operation on the boundary of the surname's core population. If the place, or places, occurs within the boundary then the surname is considered toponymic, if they are outside of the boundary the surname cores are listed as "not close to the place name". Although a reliable indicator of the validity of the KDE procedure for toponymic names, it is not a definitive one: in many cases the place names that many surnames were derived from no longer exist, or changes in spelling of either the surname or the place name prevent any direct match with the toponym database.

Moreover, it is unlikely that name bearers who do not share a common ancestor hail from a precise common location, but rather that that this indicates a general area to an unknowable level of precision.

The kernel density methodology was implemented in the R Program for Statistical Computing and Graphics (R Development Core Team 2011). The results were output as a text file (step 9, Figure 4-9) containing the metrics shown in Table 4-1. In addition a shapefile containing the boundaries for every surname core was also exported for further analysis and visualisation. It took approximately 113 hours to process the 29,430 surnames for the 2001 data and fewer than 50 hours to process the 19,993 surnames from the 1881 Census with a frequency greater than 100. These

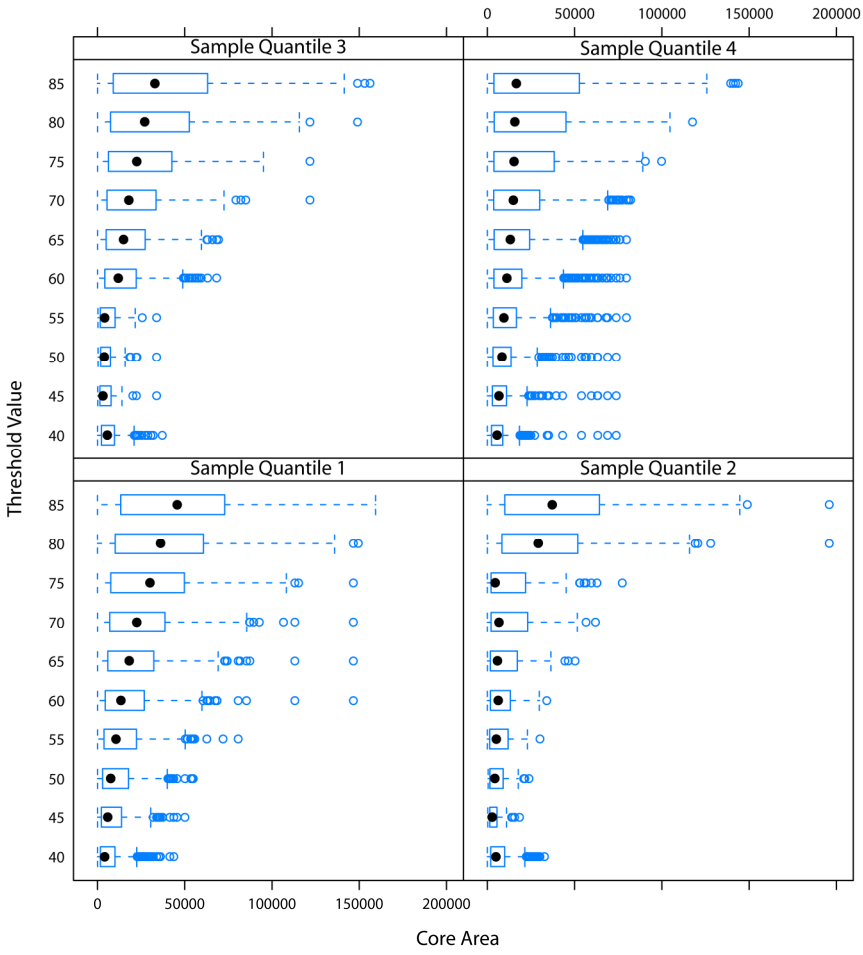


Figure 4-13: Box and whisker plots showing the effect of the threshold value on the total area of the surnames' core areas. The plots have been split into each quartile of the validation dataset to illustrate how the threshold value's influence varies with surname frequency.

computation times reach the practical limits of a desktop computing environment and would rise excessively if the surname frequency threshold were to be reduced. If this were necessary then some additional work would be required to enable parallel computation of the methodology to substantially reduce the processing time.

4.2 RESULTS AND DISCUSSION

This section will outline the results from the methodology implemented in Figure 4-9. It will focus on the 2001 results as these include a number of interesting insights into contemporary migrant surnames. The 1881 results will be also be discussed in Section 6.1.2.

Using the methodology outlined above, core surname regions were established for 92% (27,020) of the surnames analysed from the 2001 data and for 99% (19,810) of the 1881 surnames. Surnames failed to be classified as having a core area of concentration most frequently because, as Figure 4-14 demonstrates, they had

Figure 4-14: A sample of surnames classified as lacking a core area. It is clear, perhaps with the exception of Khushi, from the spatial distribution of their occurrences that it would be inappropriate to suggest areas of core concentrations in these circumstances. The size of circle reflects surname frequency. All circles are slightly transparent to show over-plotting. Published in Cheshire and Longley (2011).

extremely dispersed spatial distributions that produced very uniform density surfaces.

For ease of visualisation of the large number of contours produced by this analysis only the centroids of the core areas (as defined by the 55% population threshold) that were identified for 2001 have been plotted in Figure 4-15 to show their distribution across Britain. The map reveals a higher density of core areas around urban centres, despite the weighting diminishing their influence in the KDE. It is thought that this is largely an artefact of the clustering of migrant surnames both from abroad and also, to a lesser extent, from the internal migrations of the 19th Century (other possible explanations are provided in Section 7.3 below). Figure 4-15 (D) shows the distributions of place names that fall within a surname's area of origin, as defined by the 95% contour, that share the same spelling. The place names were taken from the comprehensive OS Placenames Gazetteer.

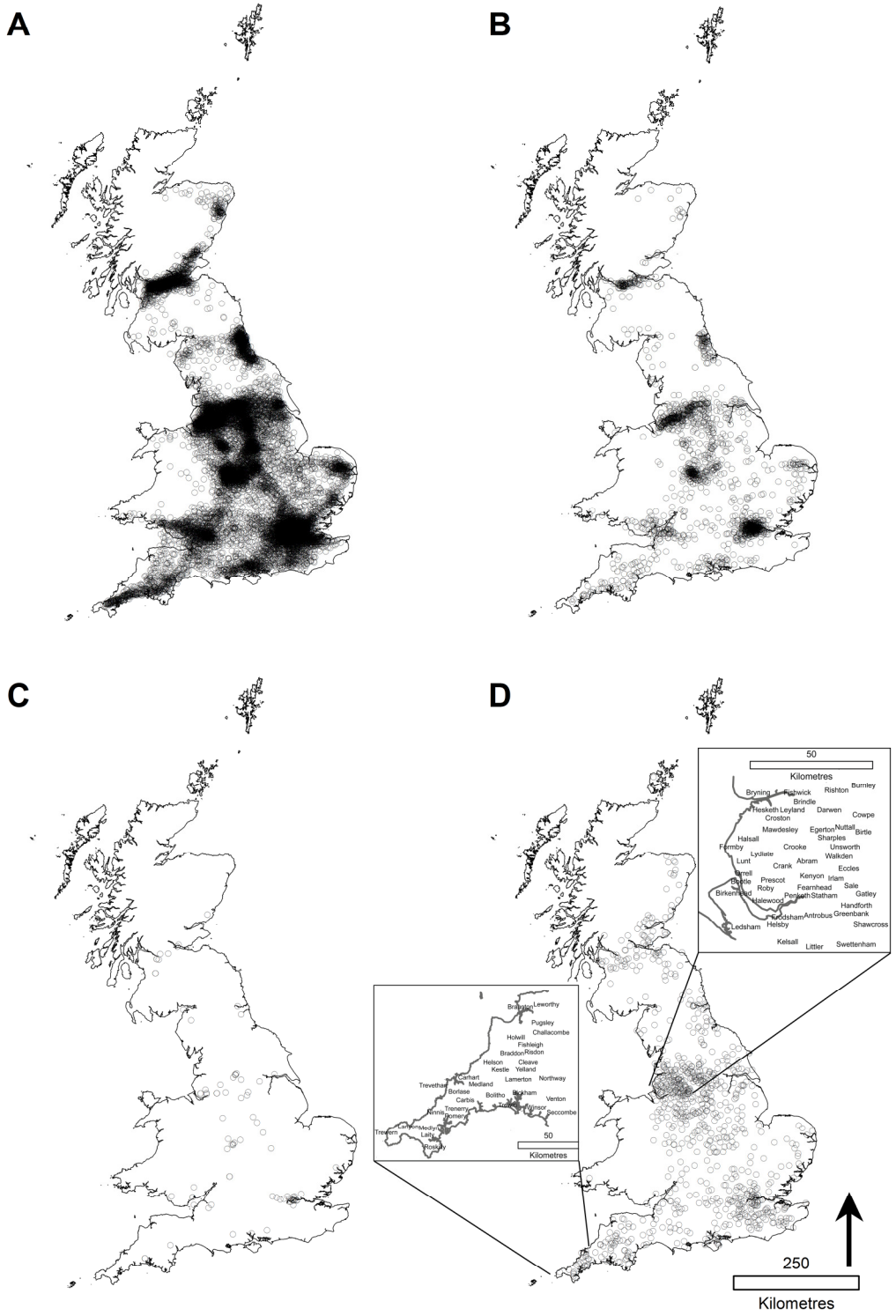


Figure 4-15: The centroid locations for surnames classified as having (A) single, (B) double or (C) triple points of origin. (D) shows the distribution of place names that occur within the core area of a surname with the same spelling, therefore indicating a high chance of a toponymic origin for the surname. The two inset maps provide the names and locations for a selection of these. Published in Cheshire and Longley (2011). See http://spatial.ly/iQhnEp for digital version.

There are two outputs from the analysis for each of the surnames: a table containing metrics and statistics (see Table 4-1); and a shapefile of the core area(s). The former may be queried for surnames with desired spatial attributes, while the latter provides interesting contextual information and facilitates visualisation and validation of the cores. The usefulness of this information will vary between applications.

There are a number of circumstances in which researchers may wish to target individuals who have historic associations with a particular location. In such circumstances, the relative concentration value and the percentage of the surname's occurrences within its core are two of the most important considerations. As can be seen in Figure 4-16, there is a clear relationship between these variables and the national frequency of any given surname. Surnames with larger populations tend to

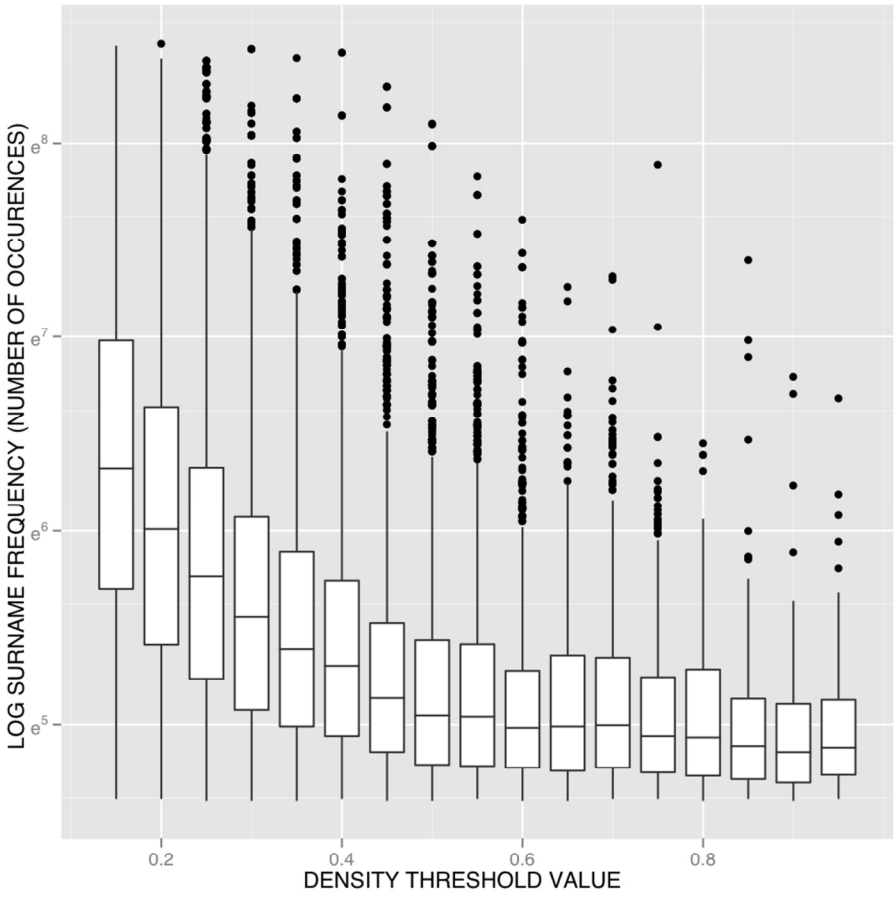


Figure 4-16: A box and whisker plot illustrating the impact of surname frequency (X-axis) on the density value (Y-axis) required to create a contour that encapsulates 55% of the surname's occurrences. For ease of plotting, only surnames with a frequency of less than 5,000 have been included. Published in Cheshire and Longley (2011).

have lower density values, because most common surnames are more evenly dispersed across Great Britain. Of particular interest are the surnames that are higher than the mean values at each density value shown in Figure 4-16. These are surnames with larger populations than many others in the same relative concentration band that are the most highly concentrated. The greater a surname's frequency within the 55% threshold the more likely it is that the surname was close to exceeding the threshold value when the contour was drawn around higher density values on the surface.

Figure 4-17 shows four examples of the outputs produced by the analysis. Figure 4-17A shows the surname "Bamber". The distribution is tightly clustered in a single area of (north-west) Britain. The contour line therefore represents this closely with a tight match to the boundary of the cluster of points. This figure also shows the cluster of points in London that are common to most surnames but that are the outcome of the sheer population size of the conurbation.

Figure 4-17B presents the somewhat different case of the surname "Palin". Here, the point distribution is also clustered in the North West but less tightly than that of Bamber. The contour thus includes more empty space in order to capture 55% of all occurrences. This can be deduced by observing that the density of that name (frequency/ core area) within the core is only 0.10 per km^2 compared with 0.38 Bambers per km^2.

The surname "Khalil" (Figure 4-17C) demonstrates a multicentred pattern that is characteristic of many names recently imported from abroad, with tight clusters centred upon populous urban areas. This has been captured by the three contour lines. In many other cases the surname is confined to a single core around Greater London. As immigrant surnames become established, usually in London in the first instance, so their core areas typically expand into metropolitan suburbs and non-contiguous locations lower down the settlement hierarchy.

Figure 4-17D illustrates the problems of grouping surnames together that have similar spellings. It is acknowledged that many surnames were derived from the same word, or represent subtle variations in spelling from the same "root" surname and combining such surnames is common practice amongst genealogists (Hey 2000). With an automated approach it is challenging to differentiate between two surnames with very similar spellings that are unrelated to each other and those that are from the same origin but spelled differently. If two surnames are variants of a common root spelling then it is likely that they will share very similar spatial distributions (Manni *et al.* 2004), and that this will be reflected in the classification. Using the approach outlined in Figure 4-9 it is clear from Figure 4-17D that Sharples and Sharpless have very different spatial distributions, with the former confined to the

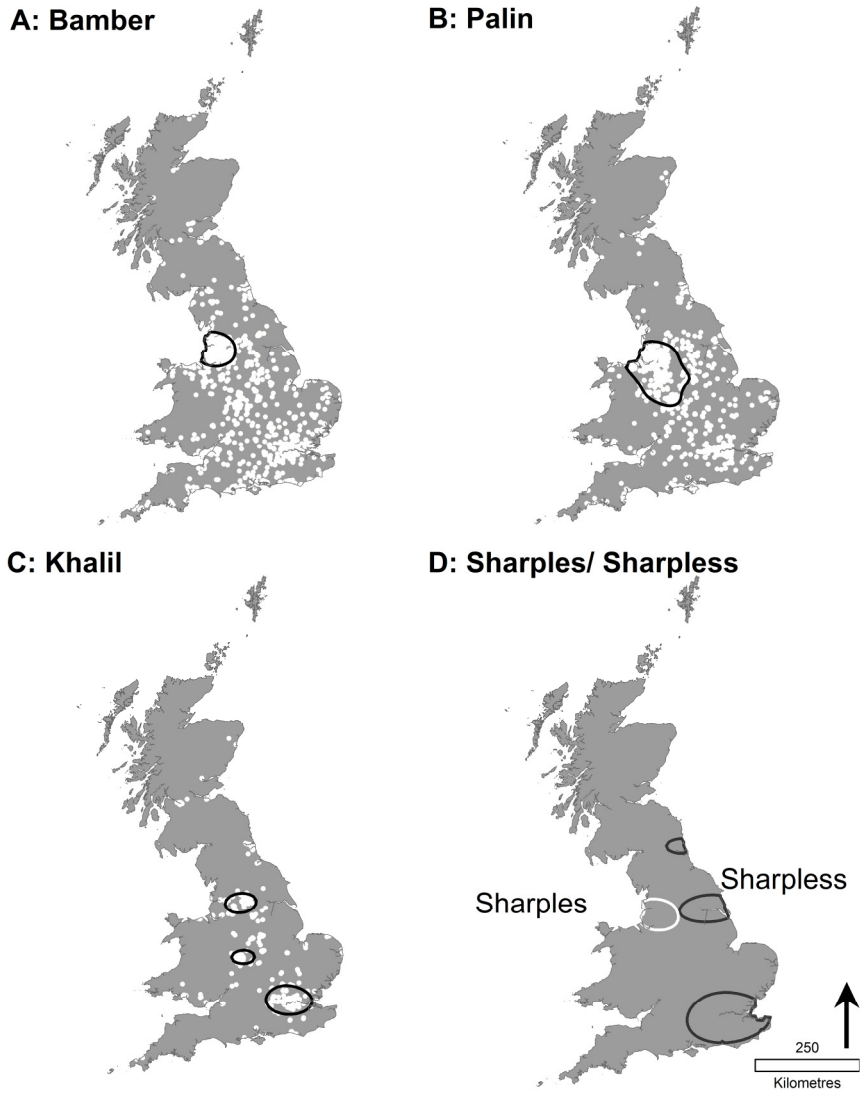


Figure 4-17: A, B and C provide examples of surname cores with their underlying point distributions. D (intentionally without points) is designed to show that surnames with similar spellings can have very different spatial distributions. Published in Cheshire and Longley (2011).

northwest and the latter to three areas on the eastern edge of England. Had these two surnames been combined the resulting spatial distribution would have clearly been misleading.

Figure 4-18 demonstrates the importance of using an appropriate population threshold to define a surname's core area and a density threshold to define its area(s) of origin. There is a fivefold increase in the number of surnames classified as having

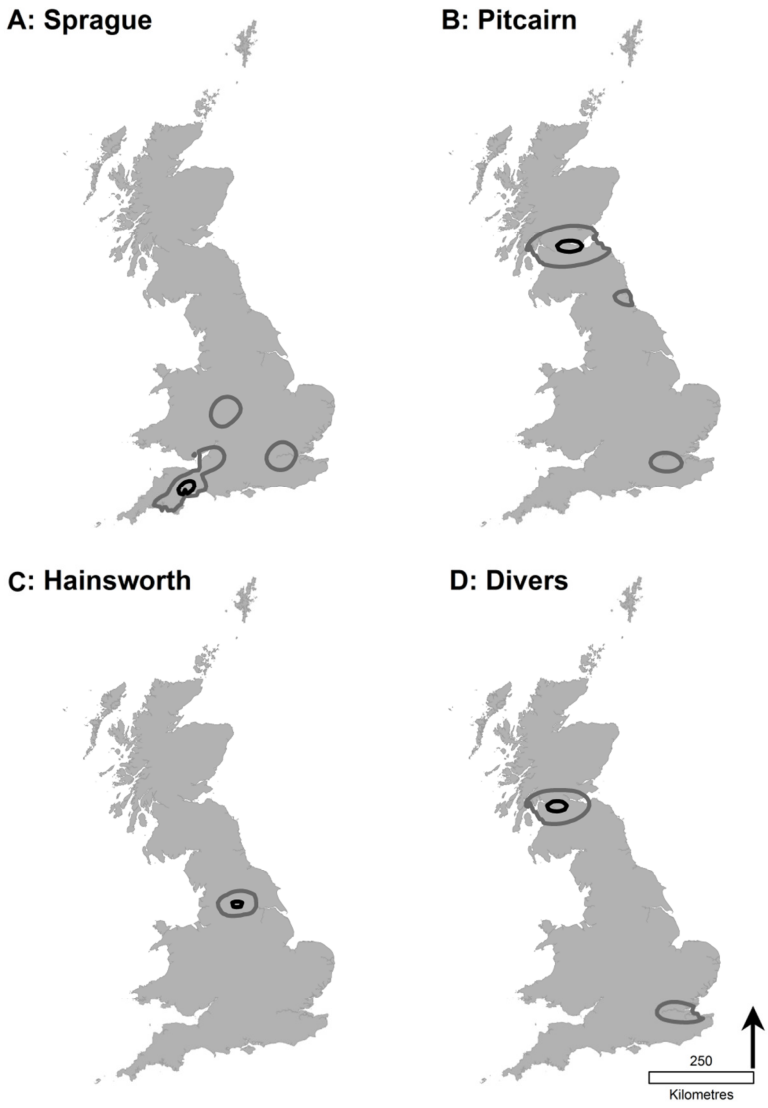


Figure 4-18: A comparison of areas of origin as classified by a 0.95 density threshold (black contour) and the surname core areas as defined by a 55% population threshold value (grey contours). Published in Cheshire and Longley (2011).

a single core when using the 95% density threshold in comparison to the 55% population threshold. The choice of threshold will depend on the area of application: thus a geneticist using regional names as indicators of distinctive regional genetic structure would be more interested in the higher threshold than a genealogist motivated to trace members of a family tree, for example. This is illustrated in Figure 4-18A where the elongated core produced by the 55% population threshold is much more extensive than the 95% density contour, and it is the latter that provides a much better indicator of the surname's point of origin. In the case of a tightly clustered surname, such as Hainsworth (shown in Figure 4-18C), there is little difference between the shape and extent of the contour drawn at 95% on the density surface compared to that around 55% of the surname occurrences.

4.2.1 TOPONYMS

The identification of toponymic surnames using the method outlined above should not be treated as conclusive. Nevertheless this classification matched 3680 place names to surnames, of which 851 were within their namesake's core areas (based on the 55% contour). The latter have been plotted in Figure 4-15. Very few toponyms exist to the west of the Welsh border or in the Scottish Highlands. Central Scotland and its east coast contain the majority of toponyms beyond England and the two most northern toponyms are Sabiston and Foubister on Orkney. Toponyms in England are most prevalent to the North West (Inset B) with additional concentrations in the South East and Cornwall (Inset C). Based on known naming preferences, this distribution is to be expected.

There are a number of caveats, however, in the ascription of toponymic surnames to localities. The analysis was conducted on a wide selection of names using two contemporary datasources (2001 Electoral Register and OS 2010 Gazetteer) that obviously cannot entirely reflect the historical circumstances under which surnames were created. There do not appear to be any toponymic names that pertain to 'deserted villages' or similar settlements. No effort has been made to relate present day frequencies of any particular names to historic settlement sizes. A further issue

that this raises is probably ultimately unfathomable – the precision with which toponymic names can be ascribed to a specific location. Thus the name 'Rossall' is unlikely to have been ascribed to all or many of the many residents who lived there (since everyone would be called the same name) but rather to someone who had moved from the place or hailed from its general direction. It is probably not productive to speculate about the nature of the differences between toponymic affinity and location of residence, though there may be some systematic trends in differences between regions of Great Britain and between types of toponym (a point location for Chester makes more sense than one for Cheshire, for example).

4.3 TEMPORAL COMPARISONS

The increasing size and geographic mobility of the Great British population since the 19th Century is well documented (Coleman and Salt 1992) and has clearly impacted on the contemporary spatial distributions of surnames. The extent to which population mobility will degrade or alter many of the historical surname distributions maintained for generations, however, is perhaps exaggerated. Pooley and Turnbull (1998), for example, state that

"there is a strong evidence for the universal nature of the migration process with more evidence of stability than change over time and space, and between different groups of the population" (p. 330).

Our contention has been that such stability is demonstrated by the spatial distributions of surnames with their continued concentration in their area(s) of origin. The typology above has demonstrated that for a large number of surnames (reflecting a significant proportion of the population) this is the case. A more detailed analysis is provided below that compares the distributions seen in 1881 with those present in 2001. Only a fraction of the temporal comparisons that can be made using the surname typology are discussed here; in addition the emphasis is on the description of the results from a novel methodology rather than attempting to provide in depth explanations for the patterns shown. The first of the three examples explores the extent to which the spatial distributions of surnames have become less concentrated and more dispersed over the past century or so. The second example demonstrates how, despite a dominance of stability, there have been considerable shifts towards urban areas for a number of surnames and the final example focuses on the stability of surnames within an area (rather than surnames on an individual basis).

4.3.1 SURNAME DISPERSION

Population growth and the introduction of many new surnames from international migration are likely to have increased the relative dispersion of surnames in Great

Britain. Metrics derived from the 55% population threshold contour, explained above, can provide useful indicators of this. Table 4-3 contains a number of metrics that summarise the contours produced with both the 1881 and 2001 surnames. The table suggests, perhaps surprisingly, that there has been a slight decrease in the area covered by the 55% population threshold for each surname. This could, in part, reflect the finer scale input spatial units for 2001. It does also show the large proportion of surnames with multi-cored distributions in 2001 reflecting the example provided in Figure 4-17C typical of many migrant surnames. This is also enforced by the fact that the mean surname population within each core area is significantly smaller in 2001 than when compared with 1881. Despite the smaller area of each 55% contour it is clear that surnames have become more dispersed with only half the surname population density within each area.

Comparisons between 1881 and 2001 are more informative when looking at direct linkages between the same surnames in 1881 and 2001. The c.3,000 surnames used here are classified as having a single core in both years. Table 4-4 provides the same metrics as Table 4-3 but the figures conform more closely to expectations. It shows that, for example, the area required to encompass 55% of a single-cored surname's population in 1881 was significantly smaller on average than the area required today. This is accompanied by an unsurprising reduction in surname population density in

Table 4-3: Metrics to summarise the different characteristics of the 1881 and 2001 surname core areas.

Metric	1881	2001
Mean Core Area (km^2)	5787.96	5457.71
Number of cores:		
1	9281.00	5501.00
2	6310.00	6834.00
3	2701.00	6383.00
4	1195.00	4769.00
>4	323.00	3945.00
Mean Surname Pop. Within Core Area	1020.04	472.14
Mean Percentage within Core Area	61.50	59.31
Core Area Pop. Density	0.12	0.06

these areas, despite a small increase in the mean number of bearers encompassed. Figure 4-19 supports this with a much greater number of core areas skewed towards fewer than 5000 km^2.

Table 4-4: Metrics summarising the different characteristics of the 1881 and 2001 surname cores. In this case only single-cored surnames are used that were present in both years. This enables more direct comparisons to be made.

Metric	1881	2001
Mean Core Area (km^2)	4081.11	7683.28
Mean Surname Pop. Within Core Area	229.04	281.28
Mean Percentage within Core Area	63.43	57.08
Core Area Pop. Density	0.18	0.05

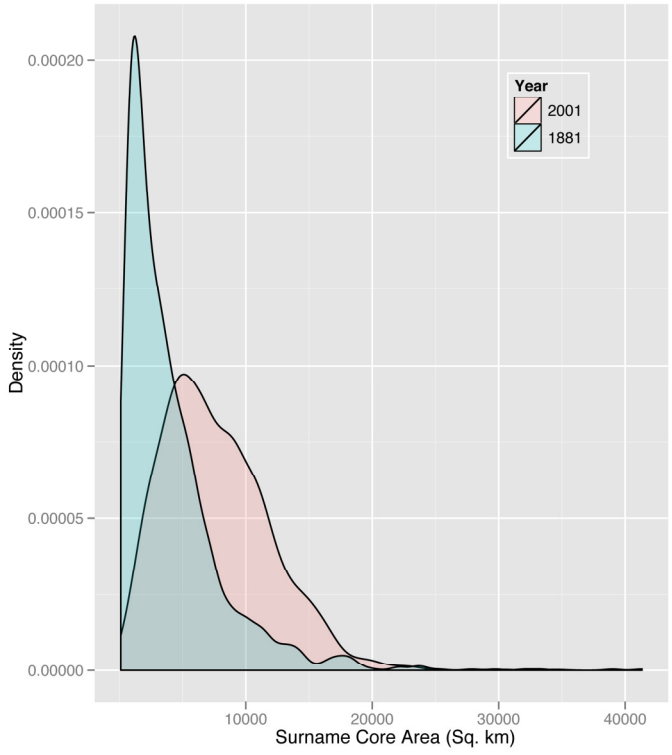


Figure 4-19: A density plot to illustrate the different distributions of core areas (km^2) between 1881 and 2001. There is a clear skew towards smaller core areas in 1881 suggesting less dispersed surnames.

Figure 4-20 illustrates the core area changes for individual surnames between 1881 and 2001. It is clear from this plot that the overwhelming trend has been an increase in the area of single-core surnames. The plot also shows that surname population appears to have little impact on the degree to which this is the case.

Overlaying the 1881 surname cores with those for 2001 in Figure 4-21 maps the increasing extent of surnames between 1881 and 2001 (the figure also shows how some surnames transition from single to multiple cores). An increase in area is not universal, with a minority of surnames actually reducing in core area between the two years. A couple of the most extreme examples ("Gairns" and "Cabot") have been highlighted in Figure 4-20 and mapped in Figure 4-22. Both surnames have relatively

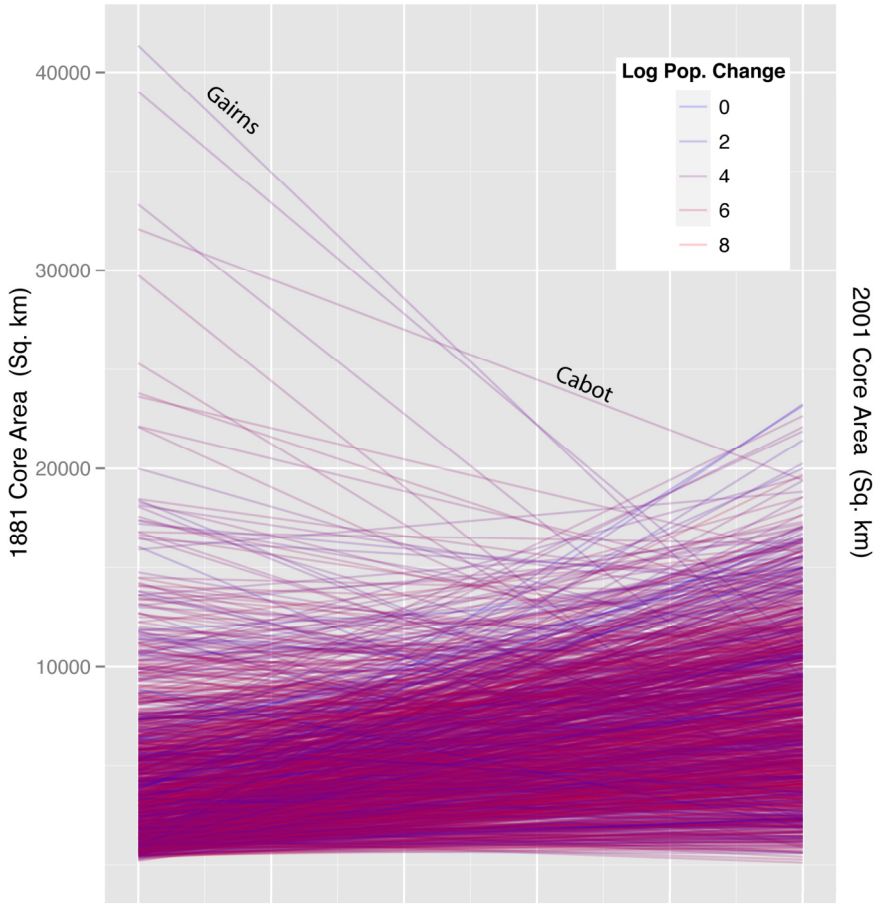


Figure 4-20: Core area changes (km^2) between 1881 (left axis) and 2001 (right axis) with two anomalous surnames highlighted. These are mapped in Figure 4-22. The plot confirms that the general trend has been an increase in the area required to capture 55% of a surname's population. The colour of the lines reflects the log of the surname population change between 1881 and 2001.

low frequencies (less than 200) and are therefore more likely to be subject to small numbers spread evenly throughout the population. In these rare cases it appears that the process has been able to create a contour from a small, dispersed, population in 1881. As Figure 4-14 demonstrates, the majority of such surnames have been classified as having no discernable core areas.

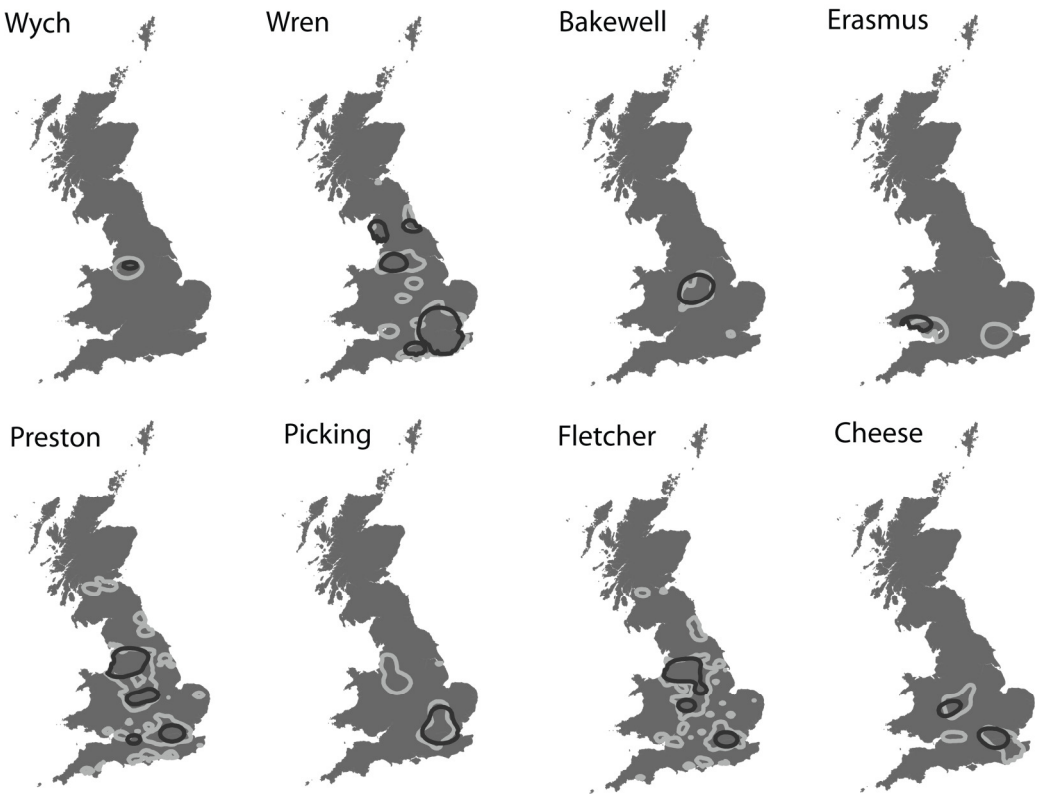


Figure 4-21: A series of maps showing the change in core areas, as defined by the 55% population contour, between 1881 (in black) and 2001 (in grey).

Figure 4-23 indicates actual percentage (above the 55% threshold) contained within the core area. As expected the dominant trend is a decrease from 1881 to 2001, suggesting that the surname population has become more dispersed in that time. This process, again, appears unrelated to the change in the surname population over time. As with the previous example, two surnames "Watmuff" and "Dunsire" have been identified and mapped in Figure 4-22. Both surnames are low frequency and continue to concentrate in the same area but have become more dispersed. Whereas the 1881 contour lines suggest a tight distribution, those in 2001 have followed a

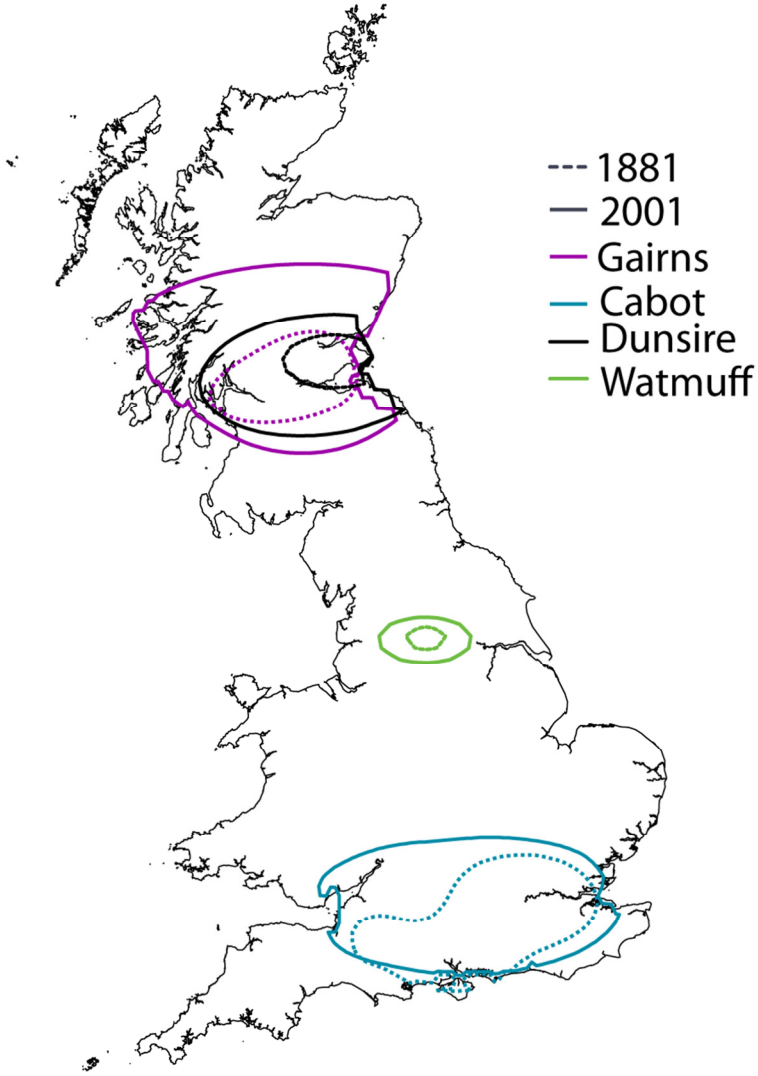


Figure 4-22: The 2001 and 1881 core areas for surnames identified in Figures 4-20 and 4-23.

more gradual density distribution that peaks in a similar location but that needs to be more extensive to account for 55% of the population. The relatively low percentage of the population contained within the contemporary cores (less than 65% in both cases) suggests even greater dispersion beyond the contour line identified.

Measures of surname dispersion are useful to gauge the integration of a surname with its surrounding population. In the case of migrant surnames, for example, one would expect increased dispersion (as measured by an increase in the size of their core area) over time. In addition, surnames that have remained within the same bounds as 1881 are likely to represent static populations of particular interest to population geneticists amongst others. Finally, temporal analysis of the dispersion of surnames is likely to provide a proxy for rates of population flux and also indicate the point in time when a surname becomes too dispersed to enable inferences to be made about its origin or past history.

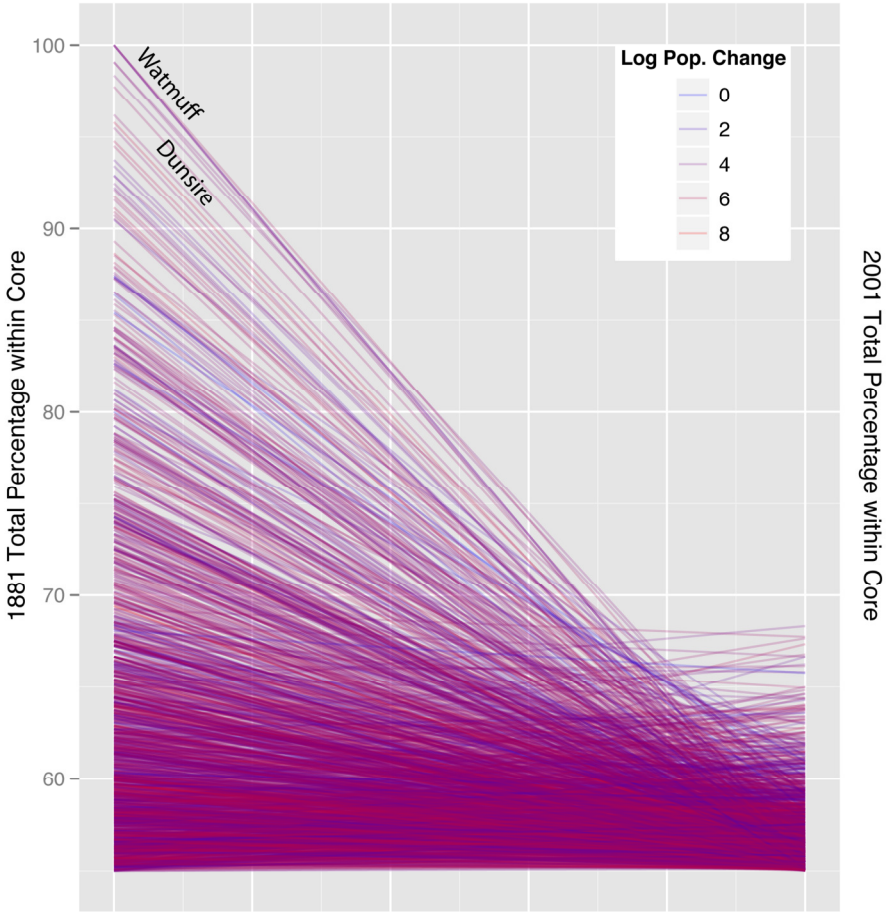


Figure 4-23: The percentage of the surname's total population contained within the core area for 1881 (left) and 2001 (right). It is a clear that there has been a dramatic reduction for many surnames, suggesting that they have become more dispersed over course of a century. The colour of the lines reflects the log of the surname population change between 1881 and 2001.

4.3.2 SURNAME MIGRATION

Interesting comparisons can be made between a surname's area(s) of highest concentration, as identified by the using the 95% density threshold for 2001 and 1881. This can be achieved by taking the mean or median coordinates of such contours to provide a centre of population for the surname that is more useful than straightforward mean or median population centres for a number of reasons. Primarily, a straightforward mean of the coordinates associated with each surname bearer (or counts of bearers), even when weighted, is easily distorted by surnames occurring at the limits of the target area. For this reason the majority of the surnames' mean population centres occurred around the geographic centre of Great Britain. Using the mean coordinates of the core contour simply provides the centroid if there is a single core area or the mean centre of a number of core areas. The former is the best representation whilst the latter is less common and will not be affected by extreme points, unless the surname had three or more widely spaced cores. Treating the centroid locations of the 1881 contour(s) as the origins and the equivalent for 2001 as the destination, a vector can be drawn between the two. This process is illustrated in Figure 4-24. It should be noted that the vectors show the shift in the centre of highest concentration of the surname, not the trajectories of entire groups of bearers. Equally it is not always the case that a surname declines in

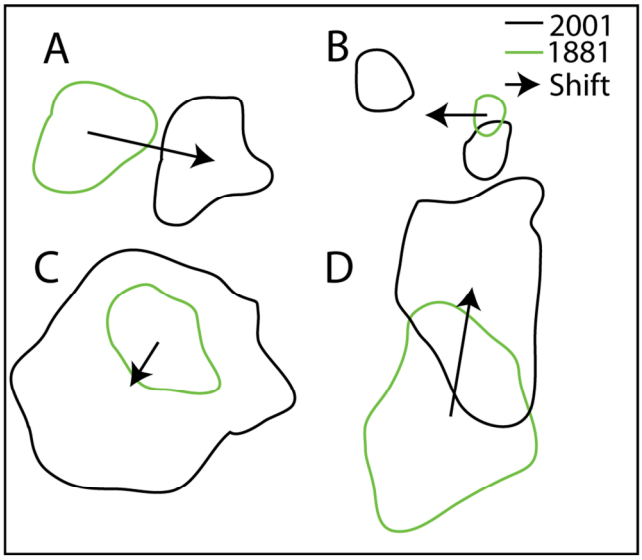


Figure 4-24: Illustrative examples of the likely scenarios that would cause a shift in the centroid of a surname's core area.

one area to increase in another. Movement may simply reflect the "diluting" of a surname due to an influx of migrants or increased concentration as a result of declines in other surnames. The vectors therefore show a shift in emphasis rather than simply presence or absence.

As Figure 4-25 demonstrates, this simple technique provides a surprisingly detailed picture that reflects known trends in Great British migration over the past century. The figure contains vectors for 15,320 surnames listed in both 1881 and 2001 and illustrates the dominance of small shifts in the core locations. It also provides an indication that the method used here is more effective than the straightforward mean population centre approach, as there is no clustering of activity around the centroid of Great Britain.

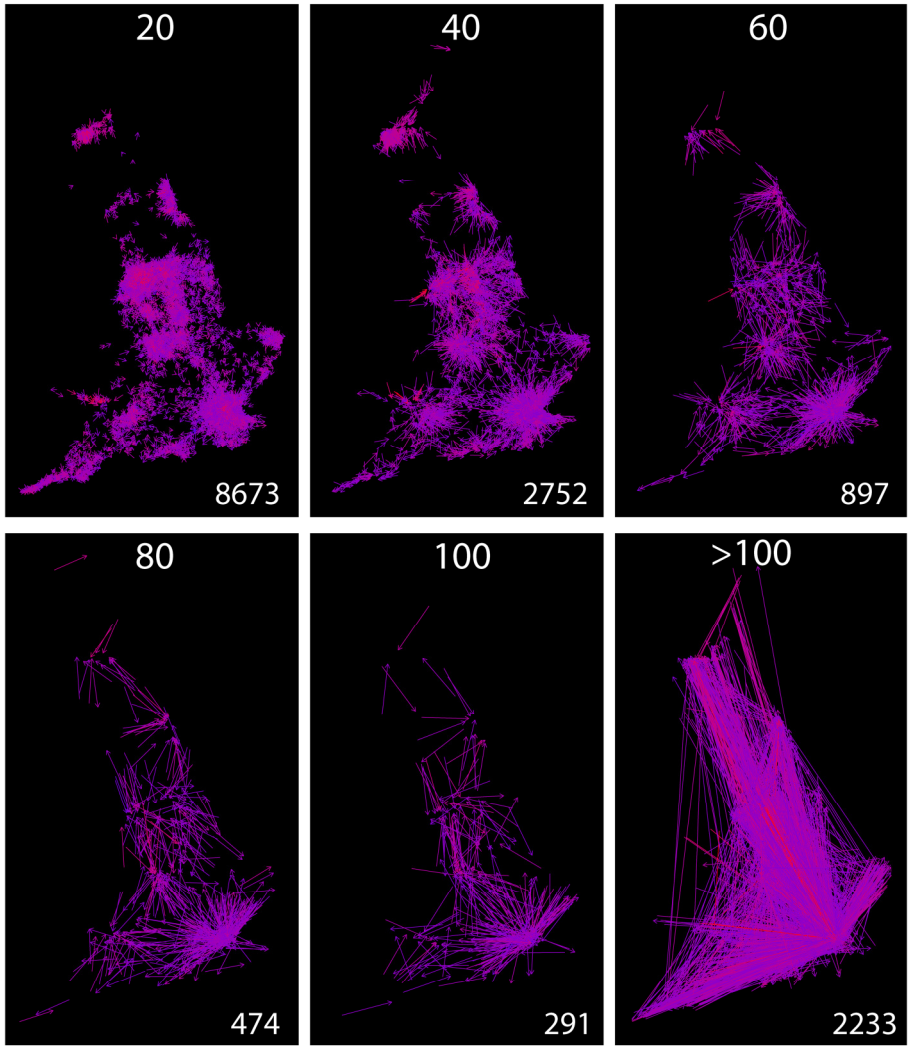


Figure 4-25: Changes in the core area centroid locations of 15,320 surnames between 1881 and 2001. Numbers along the top represent the maximum movement distances (in km) within each plot and the numbers along the bottom represent the number of surnames mapped. Redder lines represent higher 1881 surname frequency.

The vectors generally confirm the stability of British surname distributions. 75% of the surnames classified within core areas in 1881 and 2001 moved less than 40km and the majority of these were to the nearest urban area. The largest shifts (over 100km) appeared to be, unsurprisingly, between the more distant populated areas with London as a hub. The exception to this is a clear movement of Cornish surnames to the capital. It is also clear that aside from a few shifts towards the urban south, Welsh surnames are largely absent from the map. This is due to the fact that, as discussed in Chapter 2, their patronymic origins make them unlikely to have a discernable core area of concentration. It is also reassuring that many of movements concur with the dominant migrant paths mapped in Pooley and Turnbull (1998).

Figure 4-26 takes a closer look at the vectors in Figure 4-25. Figure 4-26A and B emphasises the prevalence of small-scale shifts. The smallest of these are likely to reflect inconsistences between the two years caused by differences in data quality and

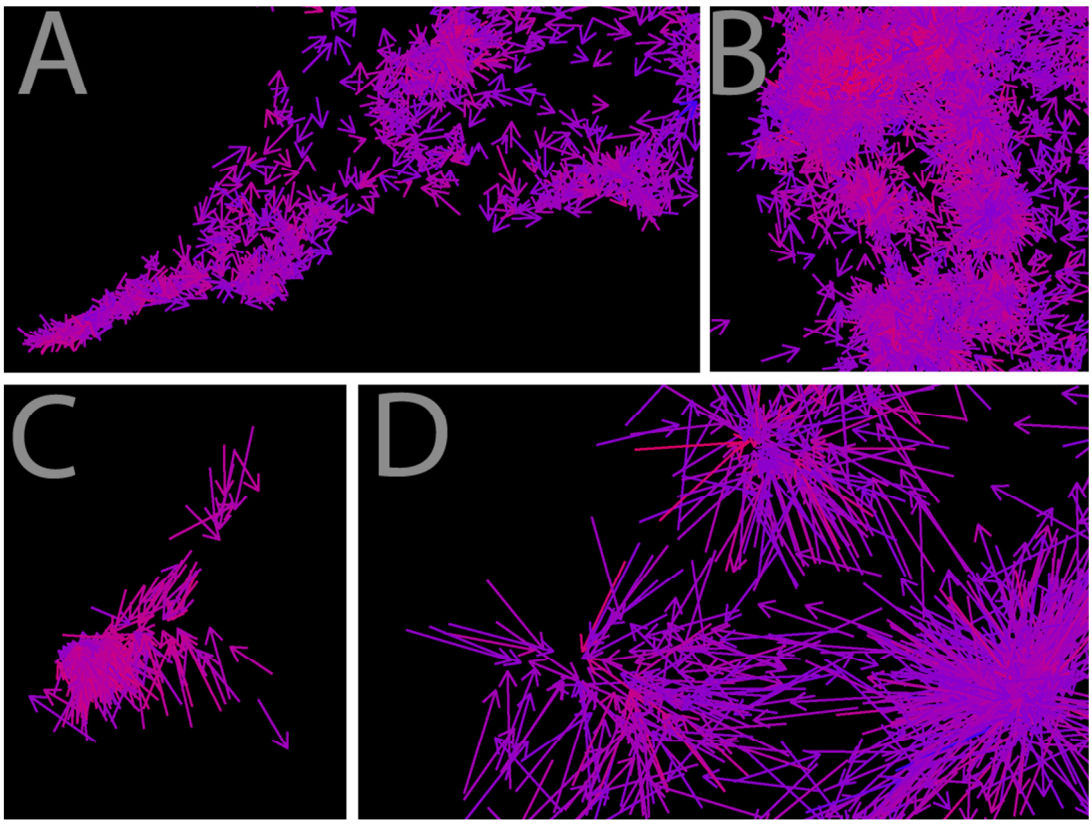


Figure 4-26: Zoomed in views of some of the vectors mapped in Figure 4-25. (A) shows moves of <20km in the southwest; (B) are moves <20km in northeast England; (C) are moves of 20-40km in central Scotland; (D) are moves of 40-60km around Bristol/Cardiff, London and Birmingham. Redder lines represent higher 1881 surname frequency.

initial levels of aggregation. It is also likely that at this scale, a finer grid would serve to reduce many of these distances. Aside from this consideration there appears to be a clear directionality to the vectors as the dominant flow is towards urban areas. The magnitude of the shifts appears to reflect the size of the "destination" urban area. London clearly dominates but there are many more subtle patterns, with the smallest shifts occurring towards towns on the south coast of England (Figure 4-26A) with increasingly long-distance flows towards the urban belt of Scotland (Figure 4-26C) and to major cities such as Bristol/ Cardiff, Birmingham and London (Figure 4-26D).

Physical constraints limit the distance and direction some surnames can travel and provide added potential for others, such as those from Cornwall, shifting to London. On this basis it is unlikely for a surname in the southeast corner of England, for example, to move more than 40km to London. Figure 4-26B (taken from the 20km map in Figure 4-25) appears to suggest clear spheres of influence for urban areas (in this case Liverpool, Manchester, Sheffield, Birmingham, Coventry) with surnames shifting towards those closest to the centre of their 1881 core concentration. Moves away from urban areas do occur and these are believed to be the result of migrant surnames sufficiently diluting those concentrating in 1881 to cause an apparent movement towards the secondary areas of concentration. Such shifts are more likely for the less common surnames.

These maps are particularly insightful in the context of inferring surname origins from contemporary distributions. On the basis that the majority of movements have been under 40km in the last century suggests that in the absence of additional contextual information, such as whether the surname is toponymic (see Section 4.2.1), contemporary distributions indicate approximate origins no more than 75% of the time. Unfortunately no comprehensive data exist for the 17th and 18th Centuries to determine the proportion of surnames that moved prior to 1881. The results produced from this analysis suggest relatively few, however, as many more surnames in 1881 have been classified with cores away from urban areas. For the purpose of assigning origin it therefore follows that the oldest high-quality population registers available should be used (see Winney *et al.* (2011)).

4.4 CONCLUSIONS

The distinctive geographies of most British surnames today are indicators of past populations in terms of their settlement structure, migration histories and ancestral characteristics. This chapter has set out an approach for measuring the spatial concentrations and dispersions of surnames that is robust and not vulnerable to spatial outliers or choice of threshold values. The application of KDE to two comprehensive Great Britain-wide names registers represents an improvement upon previous research that has used non-automated methods, incomplete datasets and inappropriate representational forms. It serves to demonstrate the utility of intensive spatial analysis for investigating surname distributions in Great Britain and provides compelling evidence that the spatial origins and diffusion of thousands of surnames can be reliably captured and summarised through a series of simple metrics.

The classification has produced a typology of surname characteristics for both 1881 and 2001 and this has enabled historical comparisons. In addition, historical data combined with the 95% contour line provides a simple and reliable method of inferring a surname's place of origin, although the value may depend upon the application. For example, the approach as described here may be better tuned to the needs of an amateur genealogist whose interests may focus upon inclusiveness – defining the maximum extent within which to search for deceased relatives – than those of a geneticist – who might be concerned to identify the distinctive genetic characteristics retained through patrilineal linkage, and hence might wish to focus upon the core of a surname area in order to maximise the chance of tracing a common ancestor.

The nature of this study and its interest in classifying such a large volume of surnames necessitated a relatively generalised view of surnames in Great Britain. However, the greatest potential of this research may be in regional and local scale studies into the geographic distribution of particular surnames. Such studies are facilitated by the insights provided by the temporal comparisons outlined in the latter sections of this chapter. These sections demonstrated the ability of surnames to indicate rates of spatial diffusion, population movements and the impact of such

things on the population stability of an area. The preliminary results point to the increasing diffusion (in terms of spatial clustering of the same surname) and urbanisation of the Great British population over the past century.

5 AGGREGATION AND REGIONALISATION

The previous chapter sought to investigate the spatial distributions of individual surnames. By combining many thousands of surname movements between two time periods it was, however, possible in the final part of Chapter 4 to begin to characterise relative stability or change in the surname compositions of specific areas. This offered a shift in emphasis from what surnames can say about people to what they can say about places. Despite the richness of the data in comparison to previous research only a partial picture could have emerged as some of the stability and variation is contained within the "long tail" of less common surnames (screened out in the previous chapter). In order to better account for these, a different areal classification approach is taken in the following chapter to unearth many of the underlying population structures in Great Britain and Europe.

The lack of research undertaken by geographers into surname regions has been an important omission in the long tradition of regional research in geography (Zelinksy 1997). The following chapters seek to address this through quantitative comparisons of surname compositions at the subnational, national and continental scale to produce a series of inductive regionalisations. The comparisons are based on a dissimilarity matrix and the regions created from a number of approaches, ranging from barrier algorithms to multiple clustering routines. The nature and extent of these 'surname regions' will be dependent on the upon the time period in which the data were assembled, and any systematic biases that arise in the data sources used.

In the light of the above considerations, a range of data and methods are explored to maximize the robustness of the regions produced. Temporal comparisons are made using the 1881 Census and 2001 Electoral Register for Great Britain. The quality of these datasets also facilitates investigation at a number of scales and with a range of spatial units to establish the degree to which these will influence the final regionalisation of Great Britain. Systematic bias, as stated in Chapter 3, is a minor

consideration with the data for Great Britain as it represents near-complete population registers. The variety of data sources combined in Worldnames to create the European-level data outlined in Chapter 6, however, produce a more partial picture and one that is subject to inconsistent levels of uncertainty for each country. For this reason a relatively novel method, known as consensus clustering, is introduced in the next chapter that provides a number of metrics to help assess the certainty associated with the resulting regionalisation.

The purpose of the next two chapters is to identify and illustrate the commonalties that characterize surname distributions at both the national and continental level. Earlier chapters demonstrated that population geneticists have laid most of the foundations for surnames research with little assistance from geography and spatial analysis. As will become clear these fields are combined here through the focus on adequate measures of surname relatedness - or surname distance - between localities or regions (studied in population genetics) and areal classification algorithms to partition space according to such distances (studied in quantitative geography). The resulting regions are a statement of within-region similarities and between region differences.

This chapter begins by reviewing why regions are a valid conceptualisation of population data, before introducing the methods used for the analysis of the surnames data for Great Britain. The effectiveness of these methods is then discussed to lay the foundations of the following chapter that both applies, in a historical and migratory context, and extends, to the continental level, the analysis.

5.1 UTILITY OF REGIONS

As with nearly all approaches to geographical research, the value of ascribing spatial data to distinct regions has been subject to debate (see Johnston *et al.* 2005: 687-690). The view taken here, in a broad sense, is in agreement with Haggett *et al.*'s (1977) statement that regions have a central status in geography because they provide "one of the most logical and satisfactory ways of organising geographical information" (p 451). The purpose of this section is to provide some broader context to the above statement and a brief review of more recent thinking towards regional geography. What follows does not purport to be comprehensive; its purpose is simply to provide sufficient context to the approach taken in the regionalisation of British (and European) surnames. Much of the following can be seen as a continuation of the discussion of Section 2.4 concerning the special nature of spatial data, brought into focus by the regionalisation concept.

The term *region* is used in a variety of ways to denote spatial compartments of formal, functional, or perceptual significance (Murphy 1991), or more generally as distinct areas on the Earth's surface (Massey 1995). Much early work treated the process of finding regions as analogous to taxonomic classification in disciplines such as Chemistry and Biology (see Haggett *et al.* 1977). Early formalisations of the concept came from Brown and Holmes (1971) who classify regions as either functional or uniform: functional regions are composed of areas (commonly in the form of spatial units) that have more interaction with each other than with outside areas, whilst uniform regions are formed from areas in which some specified respect are homogenous (Brown and Holmes 1971). In the latter case it is often a requirement that the spatial units are contiguous. Such definitions originated from geography's Quantitative Revolution (see Section 2.4) and may disregard the historical and geographical variability of regional development or the genealogy of regional formations (MacLeod and Jones 2001). These criticisms prompted many to turn to more theoretical disciplines for insights into spatial patterns (Pudup 1988). The inevitable outcome was the assertion that regional geography provided a simplistic interpretation of the complexity of human processes (Cloke *et al.* 1991) that failed to register a deeper concern with the

"social construction of places and with experiential meanings, interpretations, and emotional repertoires of human subjects- not least those relating to their surrounding environment, sense of place, lifeworld, and attachments to their place of dwelling"

(MacLeod and Jones 2001: 673).

Accounting for these concerns, the "new regional geography" of the 1980s as outlined by Gilbert (1988) provides the following classification of regions as:

- A local response to capitalist process.
- A focus of identification.
- A medium for social interaction.

(Gilbert 1988: 209-213).

All three can be applied to surname regions to some extent. For example, the 1881 distribution of surnames will reflect movements as a result of the capitalist processes of the industrial revolution. The latter two are most applicable here as they refer to the processes, outlined earlier (Chapter 2), that contributed to surname creation.

The relevance of regional differences, however they are conceived, in modern society has been questioned on the basis that improved communications and transport links have led to the shrinking of distance and homogenisation of differences between places (see Massey and Jess 1995). One could, therefore, expect the strength of relationships within regions and the differences between them to have decreased over time. However, according to Pooley and Turnbull (1998) this suggestion is contestable on the basis that a greater knowledge of other places throughout the 19th Century (through technological developments) served to heighten awareness of the differences between them and encouraged a new sense of regional identity. In addition, and perhaps most significantly, the forces of globalisation present in the 20th Century did not affect everyone equally, but rather have produced winners as well as losers; for many an increased cultural identity has become increasingly important (Pooley and Turnbull 1998). In the UK, for example, this is demonstrated through Scottish and Welsh devolution and increasingly vocal calls for Scottish independence.

5.1.1 REGIONS IN THIS THESIS

Whilst the definition of a region has changed little, it is clear from the previous section that there have been several decades of debate in geography relating to the most appropriate conceptualisation of regional similarity and difference. The approach taken here, as was alluded to earlier, takes the "classic" view of regions as a logical and useful means of partitioning spatial data. This view, in part, reflects the fact that this interpretation is a common aspect of both economic and social policy. The purpose is to create a more informed regional geography based on a ubiquitous phenomenon that is clearly the product of the more significant aspects of day-to-day life relating to the movement of people and ideas, such as distance or topographic features, for example. By creating more insightful regional aggregations from surnames it is also hoped that these facilitate more informed and generalisable research about historical population movements and structures.

Although the outcome of the methodological approach outlined below will appear to be a series of uniform regions, they will not be devoid of functional information. Looking at the mix of surnames in each region will suggest the degree to which a particular region is incorporated into national and international movements of population (likely to be fuelled by economic engagement). It is probable that spatial units with large numbers of migrants (as defined by their surnames) will be grouped together in a resulting regionalisation and can be considered as more engaged with (inter)national processes in comparison with regions of relatively uniform Anglo-Saxon surnames.

Finally, in the light of critiques of the inductive generalisation approach (as espoused in the 1960s) that is used here, a couple of additional justifications might be added. Firstly, Pooley and Turnbull (1998) argue that the refocusing away from "mechanistic and quantitative" approaches to those better suited to identifying processes "of social and cultural change affecting both individuals and communities" has been detrimental to generalisations. This is because the atypical aspects of migration (and therefore not the central processes responsible for region building) have dominated

at the expense of "the everyday and commonplace dimensions of population movement" (Pooley and Turnbull 1998: 330). This comment is enforced by the richness of the data used here. There are few other ways to process hundreds of millions of records contained within the three datasets analysed in this thesis. The comprehensiveness of the data is unprecedented in the regionalisation of cultural attributes and, for this reason alone, provides a marked improvement on previous attempts to create a regional geography at both the national and international level.

5.2 METHODOLOGICAL DEVELOPMENT

One of the principal contributions of this thesis is the comprehensiveness of the datasets it utilizes. On this basis it is important to develop a regionalisation methodology that minimizes information loss whilst creating a meaningful classification from surnames and their frequencies across a range of scales and geographic units. To this end methods have been drawn from both population genetics and quantitative geography. The purpose of this section is to introduce the methods used to create a regionalisation of surnames in Great Britain. It also forms much of the theoretical basis for the European surname regions outlined in the next chapter.

5.2.1 COMPARING SURNAME COMPOSITIONS BETWEEN AREAS

Interest in developing a measure to quantify within or between population similarities based on surnames was initiated by George Darwin in 1875. Darwin– son of Charles and the offspring of first cousins – used surnames to estimate the probability of inbreeding by estimating the number of same-surname marriages in Britain. He calculated the expected proportion of these, based simply on surname frequency, and then ascribed the observed excess above this figure to marriages between cousins sharing surnames (Jobling 2001). Darwin was keen to establish whether proportions were higher in the upper classes at that time and found this to be the case with 4.5% for first-cousin marriages among the upper classes and ~2.25% for the general rural population (Darwin 1875). Darwin's early (slightly haphazard) approach was not improved upon or formalised until the 1960s.

Crow and Mange (1965) took Darwin's ideas and proposed a measure, called the Coefficient of Relationship by Isonymy, of the probability of relatedness between individuals based on the frequency of repetition of the same surname (Lasker 2002). In addition to applications in the study of inbreeding between marital partners or social groups, isonymy can be also be used to establish the degree of relatedness between two or more population groups at different geographic locations (Smith

2002). It is this latter regional interpretation of isonymy that has gained greater currency over the last decades. This is because it is much more effective as a general indicator rather than specific measure of genetic relatedness between individuals. The Coefficient of Relationship by Isonymy when used to investigate different populations or areas extends the idea of monophyly (sharing a single common ancestor) between two populations and is defined by Lasker (1985) as

"the probability of members of two populations or subpopulations having genes in common by descent as estimated from sharing the same surnames" (Lasker 1985:142).

Based on the premise that the likelihood of a gene being shared by first-degree relatives is one in two, Crow and Mange (1965) proposed the Coefficient of Relationship by Isonymy (R_{AB}) to be half the proportion of isonymy. In the two population (one in each spatial unit) case it is defined as

$$R_{AB} = \sum_{i} \frac{p_{iA} p_{iB}}{2} \tag{5.1}$$

where p_{iA} is the relative frequency of the *i*th surname in population *A* and p_{iB} is the relative frequency of the *i*th surname in population *B*. In many cases, especially when comparing international populations, the overlap is very small and this creates very small numbers. A more meaningful measure, the Lasker Distance (Rodriguez-Larralde *et al.* 1994) is used here. It is simply defined as:

$$L_{AB} = -\ln(2R_{AB}) \tag{5.2}$$

where $R_{AB} = (p_{iA} \times p_{iB})/2$. The inverse natural logarithm creates a measure that can be thought of as distance in surname space such that larger values between populations suggest greater differences between them (i.e., less commonality in their surnames). The Lasker Distance is treated as a distance measure as it produces a (dis)similarity matrix between a series of populations or areas. The (dis)similarity matrix is a rectangular array that records the Lasker Distance measures of the degree of similarity between each spatial unit and all others in the study. The values can be thought of as distance in 'surname space' with, for example, analysis at the Census

Area Statistics (CAS) Ward-level creating a matrix comprising 10,500 by 10,500 paired measures.

Doubts about the validity of isonymy studies are founded upon the fundamental assumptions that they entail. An implicit assumption is that in some previous generation each male had a unique (monophyletic) surname, and that all surnames were first coined in the same generation (Rogers 1991). It is agreed that this is not the case in Great Britain as surnames were acquired gradually, for a multitude of reasons, in a number of distinctive and separate sub- populations. Smith, for example, is a metonym describing an occupation found within every community. However, even if two populations with very similar surname distributions do not share common individual ancestors, they are nevertheless much more likely to be genetically related amongst themselves than with a population which has a quite different surname composition. Moreover, these caveats are of less concern in ascertaining a regional geography that does not make assumptions about genetic relatedness between population groups.

Although the Lasker Distance is the most widely used metric in studies such as this, a viable alternative was proposed by Nei (1973). His measure of genetic distance, originally intended for the study of DNA similarities between populations (Nei 1978), has been applied to surnames as Nei's Distance of Isonymy in a number of studies (such as Scapoli *et al.* 2007). The purpose here is to test a wide range of regionalisation methods and, in the case of Chapter 6, to propose an innovative set of clustering techniques across a large number of countries. On this basis it was thought best to avoid comparisons of multiple distance measures. The focus on a single widely accepted surname (dis)similarity measure- the Lasker Distance- keeps this aspect of the analysis fixed to provide direct comparisons of the clustering and representational issues. All of the results presented here can form the basis of further research into the utility of dissimilarity measures from both population genetics and demographics more widely.

5.3 REGIONALISATION

The establishment of a measure of (dis)similarity between spatial units or populations stored in a dissimilarity matrix forms the first step in the regionalisation process. Unless there are very few input spatial units the matrix will have too many dimensions (up to 10,500 in this case) for straightforward visualisation or interpretation, some kind of partitioning into groups or summary measures is therefore required in order to make the results readily interpretable. The following section will tackle a number of the historical and conceptual issues associated with this process before outlining the techniques chosen to capture Great Britain's surname regions.

5.3.1 THEORETICAL FOUNDATIONS

Throughout the 1960s, cartographic techniques dominated the discovery and visualisation of regions. These techniques were, and remain, effective for illustrating areal groupings at a glance and enabling differentiation between regional characteristics (Claval 1998). They are, however, limited to differentiating regions based on a single characteristic. Cartographic representations, such as the use of contours (as demonstrated in Chapter 4), of a particular surname's frequency are appropriate for individual surnames but much less effective when attempting to aggregate multiple distributions. The limitations of the cartographic approach, combined with a revolution in computing power, led to the development of automatic regionalisation algorithms.

Following Grigg's (1965, 1967) initial work, the classification of regions is based on two methodologies: agglomerative procedures and divisive procedures (Spence and Taylor 1970). To be effective, these methods require an assessment of the degree of similarity between observations. This is achieved by calculating measures of coefficients of association, correlation coefficients and distance measures. Of these, the most commonly used are distance measures (Lankford 1969). Distance measures utilise the Pythagoras Sum of Squares equation to calculate the distances between

points in *n*-dimensional space (Spence and Taylor 1970). The Lasker Distance (Equation 5.2) can be considered a distance measure as it produces a similarity matrix based on the coefficient of isonymy (Equation 5.1) between two populations or areas. Such methods became popular amongst many geographers and regional scientists as they offered the prospect of a classification based on numerical techniques (Johnston 1968).

Despite the emphasis on induction in classification, three subjective decisions are still required that threaten to undermine any objectivity of the resulting regions (Johnston 1968):

1. Whether to use an agglomerative of divisive procedure.
2. The agglomerative/ divisive method employed.
3. How to define group membership.

Since Johnston's article there has been over 40 years of research on which to base these decisions, but consensus is yet to be reached on identifying the number of clusters for a dataset when there is no information regarding the expected or optimal number of clusters (Vickers and Rees 2007). The existence of a number of quantitative methods to inform the decision about the number of clusters (see Gordon 1999: 60-65; Everitt 1972; Everitt *et al.* 2001) is testimony to the fact that user choice is informed by a number of "informal" measures, invoking different criteria upon which to base a decision.

Before undertaking any of the steps above, the researcher needs to assess whether discrete regions exist in the data at all. This is an important consideration that relates to the nature of the datasets, the scale at which they are viewed and the probable applications of the results. The theoretical basis to these considerations is similar to the representational issues in Section 2.4 so are not discussed in detail here. In the context of surname distributions, there is yet to be a single, formalised, approach recognised as optimal for detecting geographical patterns of surname distributions at various scales. It is evident from the previous chapters and past research that there are both clear boundaries and more gradual transitions in surname distributions. The former are conducive to a discrete classification of the type outlined by Johnston (1968) whilst gradual transitions may be better captured through a more continuous

representation. On this basis, a mixed methods approach comprising agglomerative and divisive clustering, barrier algorithms and multidimensional scaling, is taken below with the purpose of capturing the multiple characteristics of surname geography boundaries.

5.3.2 AGGLOMERATIVE PROCEDURES

Agglomerative hierarchical methods provide some widely used ways to form discrete groupings within a dataset (Everitt *et al.* 2001). They produce a series of partitions in the data, starting with *n* single-member 'clusters' and finishing with a single group containing all individuals (Everitt *et al.* 2001). Of the agglomeration procedures, clustering is the most widely used within regional research and neighbourhood classifications (Harris *et al.* 2005). The ultimate aim of cluster analysis is to produce groups of individuals in which within group variance is minimised and between group variance is maximised (McQuitty 1957). However, as suggested earlier, the potential to apply one of the following three definitions to group membership can confound the researcher when choosing a clustering algorithm (Johnston 1968).The individual to be assigned to the group should be closer to:

1. One member of the group than to any other member of another group; or
2. All members of a group than to any member of another group.

Applying the first definition, a classification would group individuals by their nearest neighbours, whilst applying the second definition they would be grouped according to a rank order process (Johnston 1968). The third suggests one of two hierarchical options:

1. Centroid replacement.
2. Assuming the distance between an individual and a group is the greatest distance between an individual and any of the individuals in the group.

5.3.2.1 Ward's Grouping Algorithm

Ward's (1963) clustering algorithm is a popular method of hierarchical agglomeration. The procedure forms hierarchical groups of mutually exclusive subsets in attribute space, each of which contains members of maximal similarity in

terms of the specified characteristics (Ward 1963). The algorithm begins by assigning the n initial number of observations to $(n - 1)$ exclusive sets by considering the union of all possible $[n(n - 1)/2]$ pairs for the functional relation that matches an objective function chosen by the investigator, and then proceeds by successive iteration (Ward 1963). As with other hierarchical classifications (see Gordon 1987), the outcome of clustering can be viewed as a dendrogram that establishes the relationship between each possible pair of observations. Each time two observations are joined, a new node is introduced with branches to the joined observations, the length of which is known as the cophenetic distance. This indicates the strength of the relationship between the observations (Kleiweg *et al.* 2004).

Hierarchical clustering was performed using a set of dissimilarities provided by the matrix of Lasker Distances, using the *hclust* function in R (R Development Core Team 2011). The distances between clusters at successive iterations were computed using the Lance–Williams dissimilarity measure (R Development Core Team 2011).

5.3.2.2 *K*-means Clustering Algorithm

K-means (MacQueen 1967) is a classification method that has been particularly successful within geodemographics (Vickers and Rees 2007; Harris *et al.* 2005). It is an iterative relocation algorithm that assigns each data point into one of *K* clusters until convergence to a local minimum of its objective function (Bação *et al.* 2005). Here the objective function is the sum of squared Euclidean distance (square error distortion or within sum of squares) between each data point and its nearest cluster centre (Bação *et al.* 2005). The algorithm requires initial seeds to be allocated, around which the clusters will form for the first iteration. Of the variety of initialisation methods available the Forgy method is the most widely used (Peña *et al.* 1999). This method selects *K* observations (seeds) from the data at random then provisionally assigns the remaining observations to the nearest seed (Peña *et al.* 1999). The stochastic nature of this approach reduces the algorithm's sensitivity to outliers (Bação *et al.* 2004); this is important to reduce the impact of anomalous districts with a large proportion of non-Anglo-Saxon surnames from migration. In subsequent iterations, each data point is considered for reallocation to other clusters based on

the within sum of squares (withinss) (Singleton and Longley 2008). Where reallocation occurs, the cluster centroids are recalculated until the withinss, is minimized or a specified number of iterations is reached (Singleton and Longley 2008).

R has an in-built function for clustering by *K*-means. The algorithm works on the principles outlined above utilising the Hartigan and Wong (1979) algorithm (R Development Core Team 2011). Unfortunately, *K*-means does not guarantee reaching a global optimum as the final groupings rely on the initial groupings (Fotheringham *et al.* 2007) around the locations of the initial seeds (Milligan 1980). It is therefore prudent to repeat the process multiple times, 10,000 in this case, and select the optimal objective function from these (de Smith *et al.* 2009). In addition to selecting the lowest within sum of squares (that is the result with the tightest clusters), the clustering results were mapped and assessed subjectively at every 100th iteration to get an idea of the levels of inconsistency between each run.

A key limitation for all discrete clustering methods (not just those outlined above) is the underlying assumption that the optimal number of clusters in the data is known beforehand, something that is rarely the case in practice (Vickers and Rees 2007). A subjective decision is required about the number of clusters to be created (Johnston 1968). In this context, the assignment of the optimal cluster number has been largely based on ease of interpretation and *a priori* substantive knowledge. In the case of the surname regions for Great Britain, sufficient prior knowledge increased the confidence of the decision based on only a small range of metrics such as withinss or a dendrogram. This decision was less straightforward with the European data (see Chapter 6) due to limited knowledge of the likely cultural and linguistic influences on the outcome; it was therefore taken based on a series of more complex metrics.

The complication of determining the optimal number of clusters in which to partition the data can be avoided through the use of other methods, such as edge detection algorithms (for example, Monmonier's Barrier Algorithm) and data reduction techniques (such as multidimensional scaling). Those outlined below have

been selected based on previous (albeit limited) applications in the context of surnames.

5.3.3 DIVISIVE PROCEDURES

The concept of abrupt changes, or barriers, in the distributions of individual surnames was explored in Section 4.1 through the identification of discontinuities in surname frequency surfaces. As will become clear, Monmonier's Barrier Algorithm (MBA) is entirely different, both in terms of implementation and input data, but has a shared purpose of identifying abrupt transitions in spatial data. For example, MBA has been used to represent geographical changes in surname structure in Italy and in the Netherlands (Manni *et al.* 2008). MBA is a divisive procedure that includes spatial contiguity in its calculations. The objective of the algorithm differs from clustering as it does not seek to establish maximum internal homogeneity when regionalizing (Monmonier 1973); instead it seeks boundaries where the differences between pairs of observations on either side are largest (Manni *et al.* 2004).

The algorithm is best applied to situations where the boundaries, or barriers, between regions are of greater interest than the areas covered by the regions themselves (Monmonier 1973, Manel *et al.* 2003). It operates on a matrix of observations that have been located on a map according to their relative geographic position (Manel e*t al.* 2003). Mapping the observations requires Delaunay triangulation, which is the quickest method of connecting a set of point observations on a map with a set of triangles that fills a two dimensional space completely (Brassel and Reif 1979). If conceived as a network topology, the observations are the vertices and the edges are the connections between them. Each edge is then assigned a distance derived from the data matrix (Manel *et al.* 2003). In this case Lasker Distance is used as the distance measure between locations. The first boundary is traced perpendicular to the edges of the network, equidistant from each pair of observations, starting from the edge with the maximum distance value and continuing until the forming boundary has reached the limits of the triangulation (that is, the edge of the map) or loops back to its origin (Manni *et al.* 2004). Where edges have the same value, the one followed

by a triangle with higher values is included in the boundary (Manni *et al.* 2004). The process is illustrated in Figure 5-1.

The difference between MBA and the clustering methods outlined above should be emphasized. Whilst the clustering helps to define the regions, MBA may inform an *explanation* of them by highlighting where strong boundaries exist between regions. In Italy, for example, MBA has identified barriers between populations based on genetic and linguistic data that match topographical barriers (Manni and Barrai 2001).

More analysis is required on the appropriateness of this work in the context of surname studies, not least because it holds much potential in identifying the effects of physical barriers (for example mountain ranges, rivers or sea masses) on the movement and mixing of populations. Discontinuities between surname compositions are familiar to geneticists (see Barbujani and Sokal 1990), who argue that similar boundaries represent barriers to gene flow between populations.

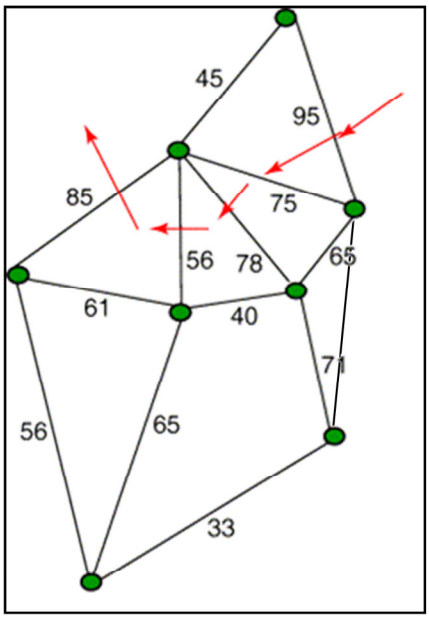


Figure 5-1: A hypothetical example of Monmonier's Barrier Algorithm. The tops of the triangles correspond to the geographic position of the observations. An example of distance between observations is illustrated by the number indicated on each edge of the triangles. In this context the algorithm obtains this value from the matrix containing the Lasker Distance between each spatial unit. The arrows represent the path of the first iteration of the algorithm. Stronger barriers between centroids can be represented with thicker lines. Source: Manel *et al.* (2003: 6).

MBA can be implemented with the standalone *Barriers* software (Manni *et al.* 2004) or the *adegenet* package for R (Jombart 2008). In this case the *adegenet* package was used.

5.3.4 REDUCING DATA DIMENSIONS

5.3.4.1 Multidimensional Scaling

In addition to the identification of discrete transitions in surname geography multidimensional scaling (MDS) is also used to show more subtle and continuous differences that depict trends or surfaces of closeness or dissimilarity between populations. Following Golledge and Rushton's (1972) pioneering work, MDS has found many spatial analysis applications (Gatrell 1981). It provides an effective summary of the degree to which regions are related to each other in 'surname space'. MDS reduces the dimensionality of a dataset from an m x m (dis)similarity matrix (where m is the number of spatial units) to an m x n matrix where $n<m$ and values of n can be treated as coordinates in relative rather than absolute space. These coordinates can be plotted to show the similarity between spatial units in this relative data space (Gatrell 1981). It belongs to the same family of data reduction methods as Principal Components Analysis (PCA). MDS can either be metric or non-metric: both variants regress distance against dissimilarities, with the former using interval measures of the dissimilarities and the latter using their rank-order. Here the metric variant is used as it is well suited to studies where the distance measures arise directly from previous analysis methods (Everitt *et al.* 2001).

MDS simplifies the data and represents them in a geographical model of three-dimensional coordinate space with Euclidean distance representing the proximities derived from the chosen measure. Each of the combinations of coordinates can be visualized in two or more dimensions to provide a visual (but not geographical) method of detecting cluster structure (Everitt *et al.* 2001). The results can be visualized as two or three dimensional scatter plots.

The MDS results can also be mapped as a more novel representation, previously used in linguistics (see, for example, www.let.rug.nl/~kleiweg/), that shows the three-dimensional MDS values on a 2-D map. The raw MDS coordinates are rescaled to values between 0 and 255 in order that they can be substituted for a value in the Red, Green, Blue (RGB) colour model. Thus each geographical unit has a unique colour assignment based on its MDS coordinates. Similar colours/ shades are produced when districts share similar MDS coordinates and therefore must be closer together in 'surname space'; likewise more colours/ shades indicate a greater surname disparity between regions. Group membership from MDS can be established by proximity of the spatial unit's the three dimensional coordinates to others or final colour allocation in the MDS maps.

For its application in this chapter and the next, Manni *et al.*'s (2004) concerns that MDS (like principal components analysis) does not provide a statistical analysis of the pattern of variation, instead portraying an interpolated landscape in geographic space, are acknowledged. Conceptually it differs little from the maps produced by Lao *et al.* (2008), or Cavalli-Sforza (2000), which rely on spatial interpolation techniques to infer genetic characteristics in areas where samples have not been taken. This, in part, is the reason why a mixed approach is adopted here by combining MDS with other types of cluster analysis in order that one set of results can provide context to the other.

In common with the previous methods, R was used to undertake the MDS calculation. To visualize the MDS colour values as a map a custom VBA script was written for ArcGIS 9.3.

5.4 ESTABLISHING SURNAME REGIONS FOR GREAT BRITAIN

There is much historical, linguistic, anecdotal and genealogical evidence for the existence of cultural and ancestral heartlands within Britain, as well as interest in them from a range of disciplinary perspectives (for example Fryer *et al.* 2004; King and Jobling 2009; Lauderdale and Kestenbaum 2000; Scapoli *et al.* 2007). However, the research into the nature of these regions has focussed on single events, serendipitous datasets, or specific regional case studies, without regard to robust measurement and comprehensive coverage across Britain. The following section provides the most comprehensive regional study of surnames in Britain undertaken in addition to the first substantive applications of the techniques to a historical data source (the 1881 Census). The results outlined below to some extent mark the contagious diffusion of names from their historic heartlands (see Section 4.3), as well as hierarchical diffusion of names imported from abroad: as such, the analysis also presents elements of a functional regionalisation of the country, particularly with respect to urban areas.

5.4.1 A NOTE ON DATA

Bearing in mind the effects of scale and geographical units, outlined in Section 2.4, two levels of geography were selected for the contemporary data analysis: Local Authority Districts and Census Area Statistics (CAS) Wards. The 410 Local Authority Districts in Great Britain have an average population of approximately 105,000. There are 8,850 CAS Wards in England and Wales, each with a minimum of 100 residents; 1222 in Scotland, each with a minimum of 50 residents. The mean CAS Ward size in the dataset is 4,300 (based on the 45.6 million observations in the enhanced Electoral Register, and 10,500 CAS Wards in Great Britain). Use of the smaller CAS Wards, although not the smallest available areal unit, was deemed the most appropriate to this research for three reasons: first, these units ensure that sufficient and consistent numbers of surnames are present within each spatial unit and hence provide a plausible basis for comparison; second, in practical terms,

computation of Lasker Distance calculations at CAS ward level for Great Britain entails creation of 9 x 10^{13} (828,131*10,455*10,455) cells of data; third, it is also a salient unit of analysis, in that most CAS Wards retain historic associations with parish areas in many parts of the country. The same calculation with smaller spatial units would take orders of magnitude longer: for example, the Lasker Distance calculation at Output Area Level would require approximately 15 days of processing and produce a matrix too large to be clustered using the methods outlined here. The geography of the 1881 Census is fixed at Registration District level (see Section 3.1.1 for more details).

5.4.2 IMPLEMENTATION

The general methodological steps undertaken here are summarised in Figures 5-2 and 5-3. The first demonstrates the steps completed to complete the Lasker Distance calculation and the second concerns the regionalisation of the resulting matrix. Initial calculation and clustering of Lasker Distances and subsequent mapping of the results produced highly fragmented patterns at the Local Authority District level of geography, in large part because of the atypical composition of surnames in the 32 London Districts that make up this unique (in the context of Britain) world city (see McElduff *et al.* 2008). Amalgamating the 32 Districts that make up London enabled the creation of a more plausible national regionalisation. In the final analysis, therefore, the Lasker distance was calculated for 379 districts. The effect of London was less disruptive at the CAS Ward level, as London Wards account for just 6% of the total number as opposed to 13% of the Local Authority Districts. The resulting patterns arising from applying from the two techniques at the Local Authority District and CAS Ward levels are outlined below.

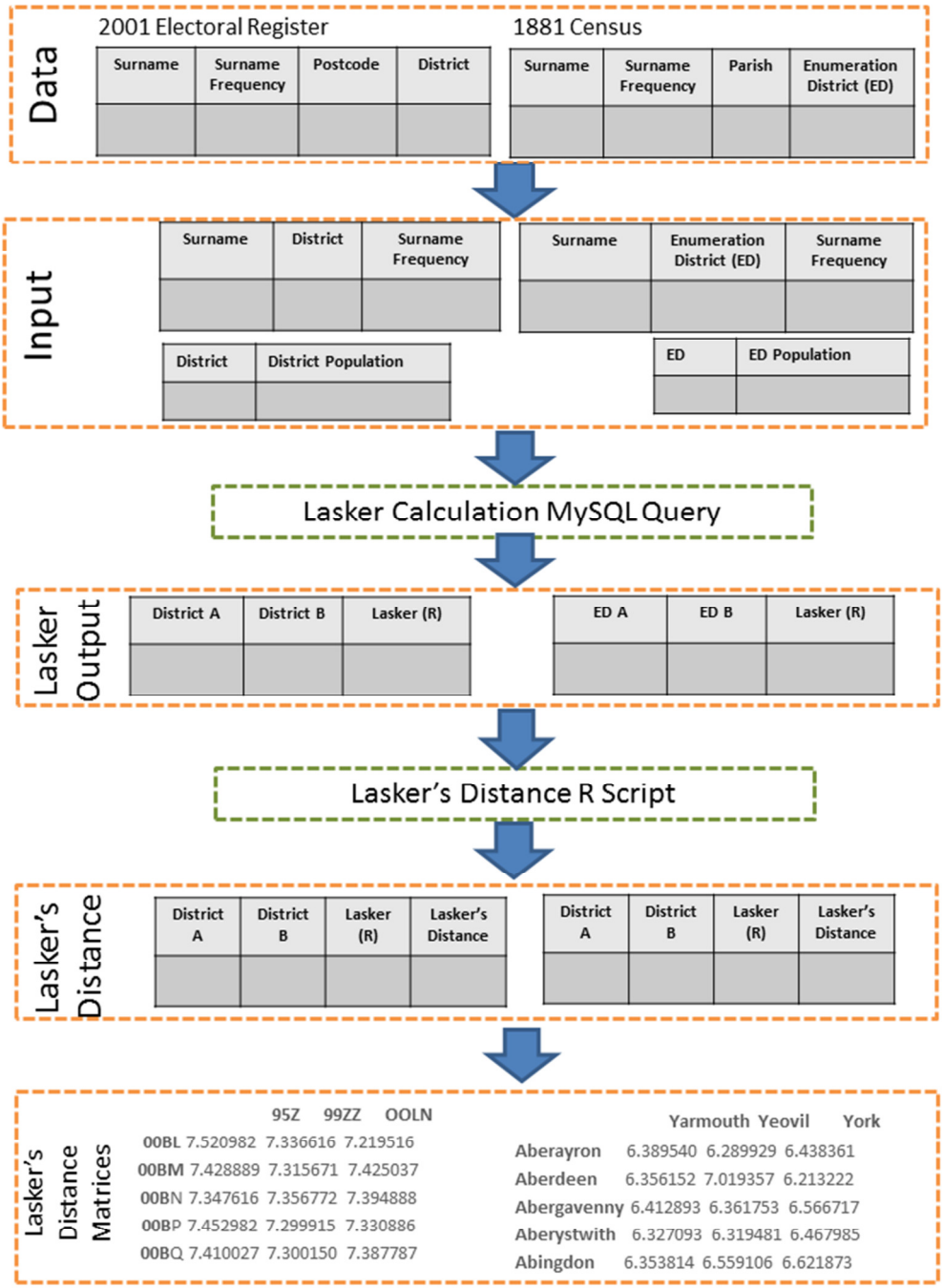


Figure 5-2: A flow chart to illustrate the Lasker Distance calculation phase of the methodology.

The large size of the datasets necessitated the use of MySQL Database software for storage and the calculation of isonymy (R_{AB}) (Equation 5.1). The 2001 dataset is significantly larger than the 1881 data, and required approximately 15 minutes of processing using a high-performance computer workstation to complete the R_{AB}

calculation. By far the longest computation was with the Ward level geography and this took approximately 12 hours of processing to complete.

The database query produced a table for each of the two time periods, with the R_{AB} values (see Equation 5.1) comparing each district with every other district in Britain (ie. a matrix of 658 by 658 (for 1881), 379 by 379 (2001 Local Authority District level), 10,500 by 10,500 (2001 CAS Ward level)). The matrices were exported from the database as .csv files and loaded as an object in the R for the Lasker Distance calculation and subsequent analysis. The resulting visualisations were produced using a combination of R and ArcGIS 10.

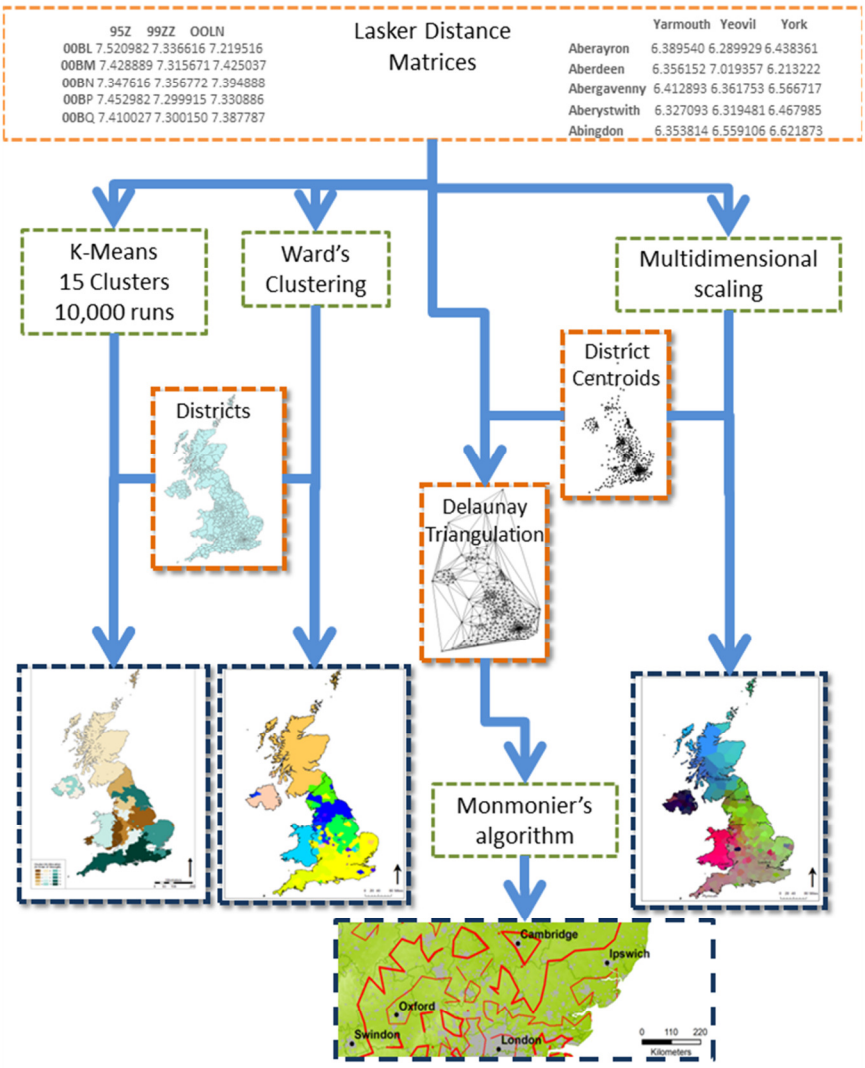


Figure 5-3: A flow chart outlining the regionalisation and visualisation phases of the methodology.

5.4.3 RESULTS

Table 5-1 shows a number of descriptive statistics generated from the paired Lasker Distance measures between the spatial units of both years. The mean distance suggests little change in the national picture of similarity or difference in surname compositions both between years and spatial units. This is slightly misleading on the basis that there has been a dramatic reduction in the range of values and an associated decrease in the standard deviation between 1881 and 2001 (ignoring the CAS Ward level results). Whether this result, plotted in Figure 5-4, is logically explained by a relative homogenisation of the 2001 population caused by the migratory processes outlined in Chapter 4.3, or simply the effect of the reduced number of spatial units is not clear. The effect of distance on surname mixing is evident in both years with the greatest Lasker Distance values recorded between Eastern Wales (Aberayron) and North Yorkshire (Reeth) in 1881 and between Slough and Orkney in 2001. The CAS Ward level statistics for 2001 offer insights into the magnitude of the fine scale variation discernable at smaller scales with its much larger range of values and higher standard deviation than the same Local Authority District level analysis.

The methods outlined above have had varying success in producing a coherent regional geography of Great Britain's surnames. The first aspect of this section relates to the general trends and historical comparisons produced from the 1881 Census and 2001 Local Authority District level Electoral Register. The second aspect will focus on the detailed analysis conducted using the CAS Ward level data for 2001 only.

Table 5-1: A series of descriptive statistics produced from the Lasker Distance calculation.

Statistic	1881	2001 District	2001 Ward
Mean	6.27	6.31	6.37
Range	0.57-10.82	2.60-7.10	1.76-11.57
Standard Deviation	0.60	0.31	0.46

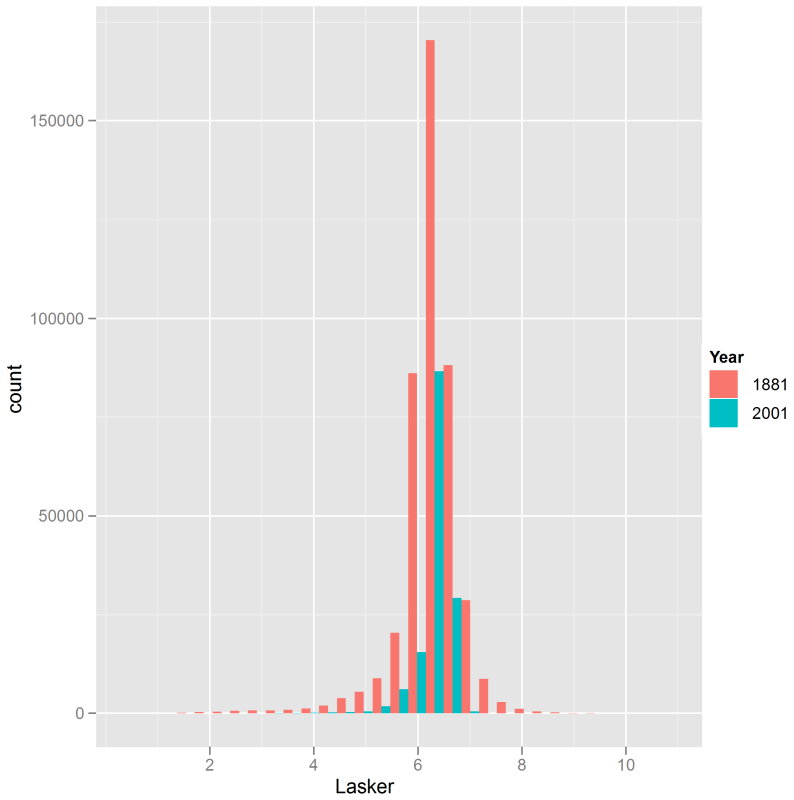


Figure 5-4: The distribution of Lasker Distance values for the 1881 Registration Districts (red) and the 2001 Local Authority District level (blue).

5.4.3.1 Monmonier's Algorithm

The barriers resulting from MBA, shown in Figures 5-5 and 5-6, present a complex picture. One of the most noticeable differences between the datasets is the concentration of barriers around London and the South in 2001, compared with a more even spread in 1881. Commonalities in the results include the Scottish border region, especially prominent in 2001, and a barrier delineating South West England.

5.4.3.1.1 1881 Barriers

In the Southwest there are 3 major barriers: one splitting it from the rest of England starting from North Somerset and going South around Poole and the second barrier tracking some way along the Devon/Cornwall border before heading East and stopping short of Exeter. A third barrier excludes Plymouth from the rest of the Southwest. In the far north a barrier splits north and south Scotland, whilst in England a strong barrier forms between the City of Durham and the rest of its county, in addition to the division between the north-eastern coastal towns and Newcastle-upon-Tyne. Moving south, Greater Manchester appears to have a number of barriers surrounding it, suggesting a number of differences between the urban area

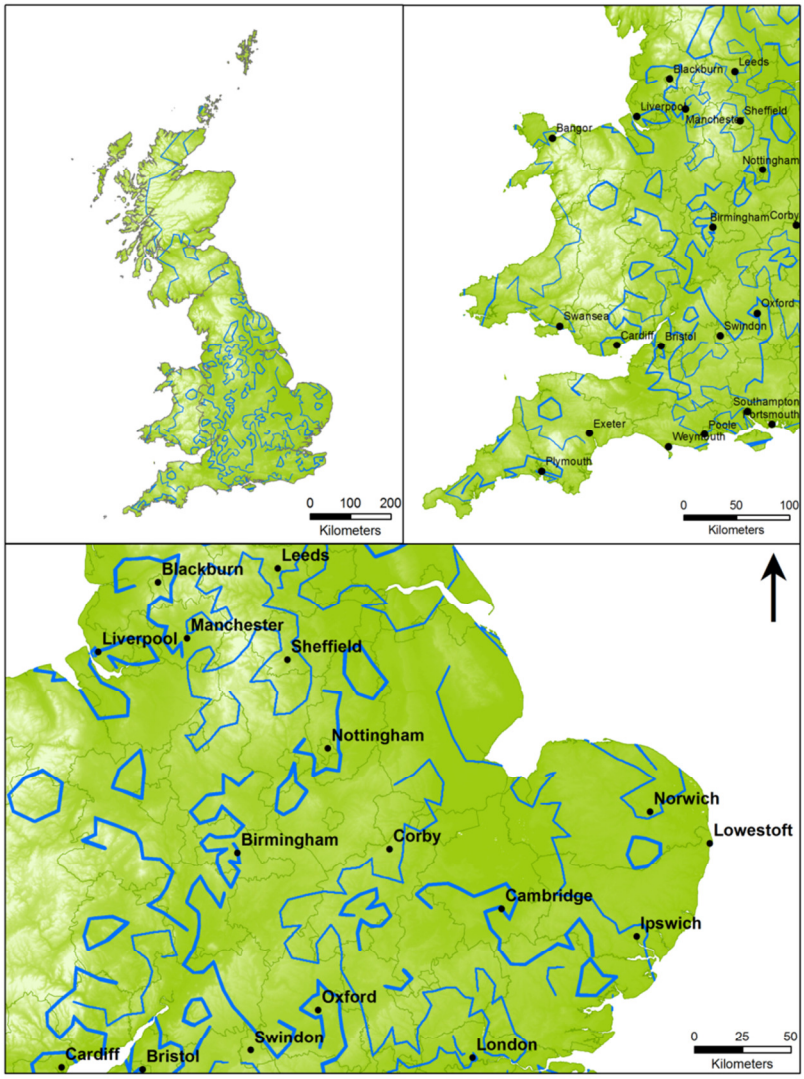


Figure 5-5: 1881 surname barriers created using the Monmonier algorithm mapped without the underlying Delaunay Triangulation and overlain on Shuttle Radar Topography Mission (SRTM) data. Contemporary county boundaries are shown in dark green.

and its more rural outskirts in 1881. In Wales there is agreement with the K-means results for 1881 as the large settlements along the south coast (Cardiff, Swansea, Newport) and Pembrokeshire have barriers differentiating them from the rest of Wales. The islands of Sheppey in Kent and Anglesey in Wales have weaker barriers that nevertheless separate them from the rest of mainland Britain. In addition many rural areas have had barriers drawn around them. This could be due to a lack of social mixing or data artefacts. Finally, unlike today, London and its suburbs do not appear differentiated from surrounding areas as the city has relatively few boundaries around it. One barrier extends through Central London, roughly following the Thames, suggesting a north/ south split in the population composition of the areas.

5.4.3.1.2 2001 Barriers

Barriers derived from the 2001 Local Authority District level data suggest an east west division in England. A barrier originates between Manchester and Blackburn tracks south, west of the Peak District, east of Derby and West of Leicester. Using this barrier Liverpool, Manchester, Stoke-On-Tent and Birmingham can be classified as western cities, whilst York, Leeds, Sheffield and Leicester are eastern cities. A barrier further south continues the east/ west split with Oxford and Basingstoke to the east and Swindon and Andover to the west.

On a regional, rather than national, scale other interesting barriers exist. For example, two barriers between Nottingham and Derby in 2001 imply a change in surname structure. In Northamptonshire, Corby is a town that has been isolated from other areas by a barrier; this division is supported by the other methods utilized in this study. Elsewhere, the coastal fringe of East Anglia creates a strong barrier from the rest of Eastern England. This suggests that the towns of Ipswich, Lowestoft and Great Yarmouth share more commonalities with each other they do with the city of Norwich. Cambridge also appears an isolated city in Eastern England with a strong barrier along its perimeter. The final barrier of interest is that which divides the region Dorset (including Bournemouth) into the more urban and coastal South East (including towns such as Weymouth, Poole and Bournemouth) and the more rural,

inland North West of the county. The northern edge of this barrier closely follows the Dorset/ Somerset border.

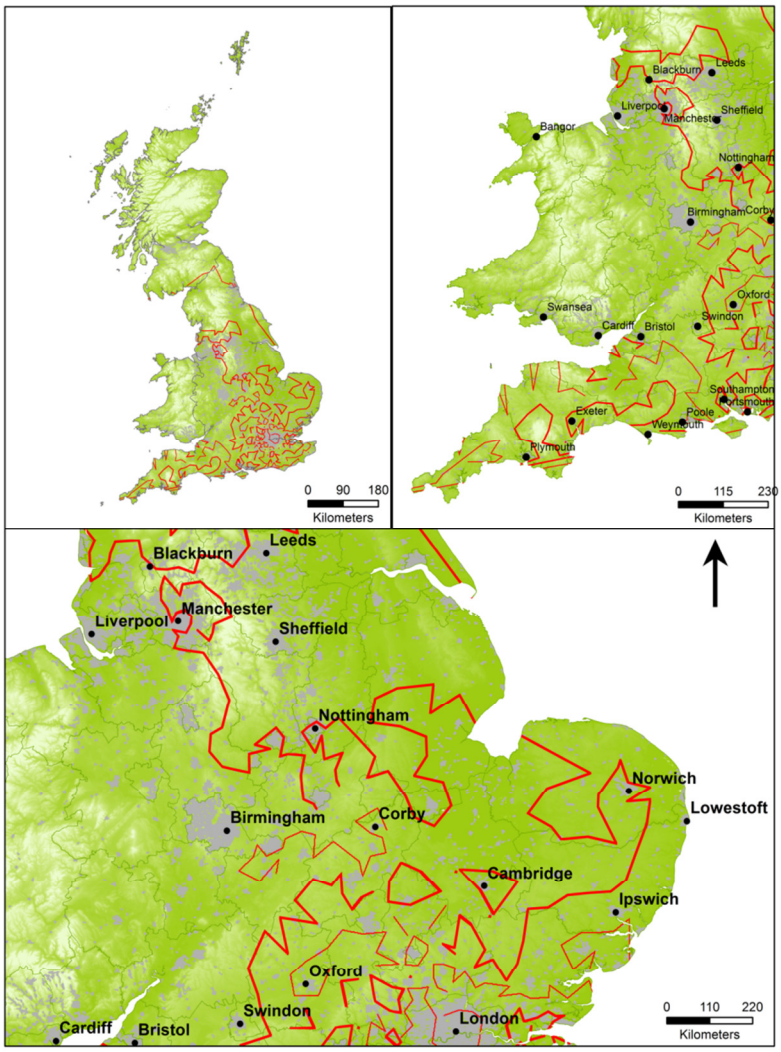


Figure 5-6: 2001 surname barriers created using the Monmonier algorithm mapped without the underlying Delaunay Triangulation and overlain on SRTM data. In addition large settlement footprints are mapped in grey and county boundaries in dark green to add additional context.

MBA appears to highlight a number of plausible barriers to surname flow in both years, but it also produces a noisy national picture. Its results are heavily influenced by a number of factors associated with the initial conditions in the data. The first is the definition of contiguity applied. Delaunay triangulation is the standard implementation but there are a number of other alternatives such as Gabriel graphs or distance based measures of contiguity that would influence the possibility of

barriers being drawn between two or more areas (Bivand *et al.* 2008). A further point is that the algorithm can get stuck in loops around spatial units significantly different from their surroundings. Such behaviour is important for identification purposes but it reduces the chance of barriers forming elsewhere. This appears to have been the case with urban areas in the 2001 output. In addition the results from MBA are greatly influenced by the spatial units and level of aggregation used. Fewer spatial units will increase the chances of barriers being created. This is a problem applicable to many aspects of spatial analysis (see Section 2.4), but it seems particularly noticeable in the context of MBA. Finally, the method is extremely computationally intensive and therefore not easily applied to more than a few hundred spatial units. On this basis alone, its implementation has been limited in the rest of the analysis.

5.4.3.2 *K*-means

From Figures 5-7 and 5-8 it is clear that the *K*-means clustering algorithm with 15 clusters produces smaller, more fragmented, regions when compared with the Ward's hierarchical clustering (shown in the next section). The procedure appears to identify groupings that are more sensitive to variations within Scotland and Wales. Unlike the Ward's clustering algorithm, *K*-means distinguishes three regions within Wales. In both years, the western tip of Wales (Pembrokeshire) has more in common with the Welsh border regions that extend along the Bristol Channel, including Newport and Cardiff, than central areas of the country. West of Cardiff into the County of Swansea and inland to the Welsh mountains region there are commonalities with the border regions, differentiating this area from the bulk of Wales. Finally the north west of Wales (the County of Gwynedd and Isle of Anglesey) appears as a distinctive area. The within sum of squares ('withinss') values associated with these observations suggest that the border region of Wales, Central Wales, South Wales and Pembrokeshire are more tightly clustered than the North West region of Wales; one could therefore infer that the degree of difference between this region and central Wales is less profound than between the other Welsh regions highlighted.

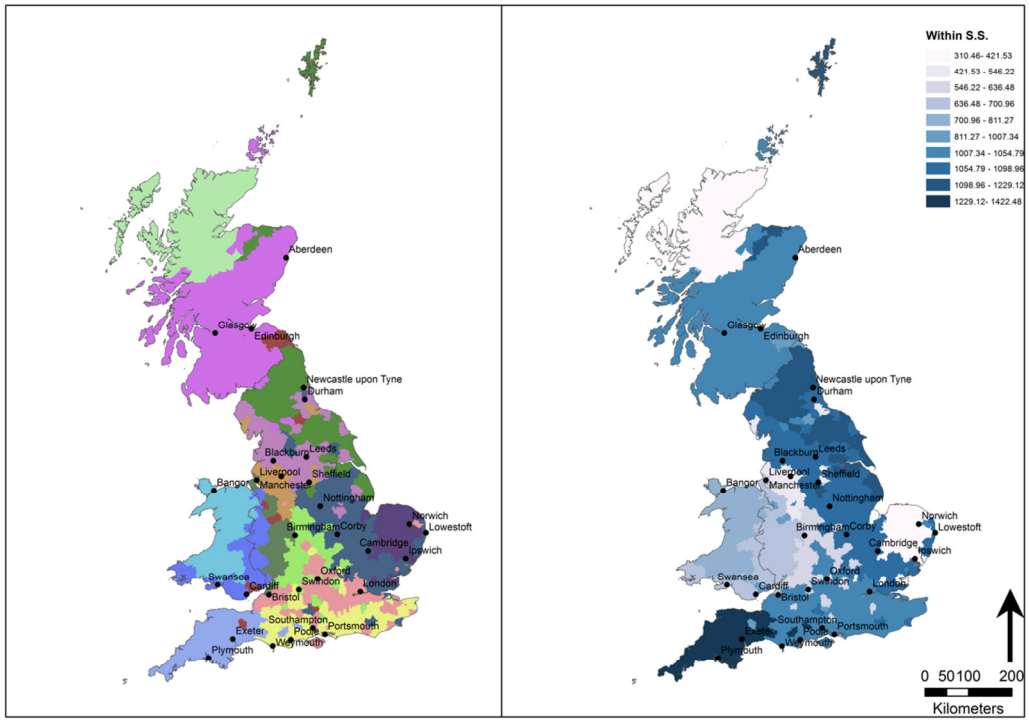


Figure 5-8: 1881 *K*-means clustering maps showing the surname regions at *K*=15. The cluster allocations (left) are represented by unique colours and lower withinss values (right) are represented with darker colours to identify tighter clusters.

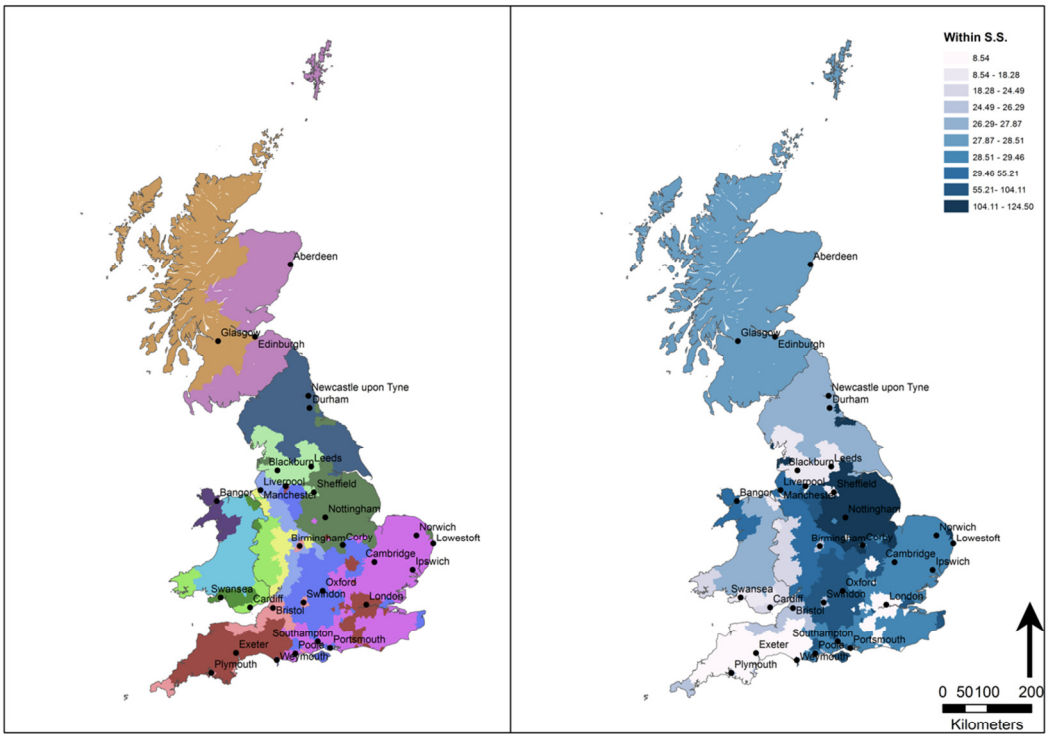


Figure 5-7: 2001 *K*-means clustering maps showing the surname regions at *K*=15. The cluster allocations (left) are represented by unique colours and lower withinss values (right) are represented with darker colours to identify tighter clusters.

In England, the three tightest clusters have created a region for the South-West of England, the North-West conurbation of Liverpool and Manchester and London with suburbs. Scotland can be approximately divided into highlands and lowlands, with the Shetland Islands sharing a greater affinity with the latter. This split is interesting as it does not appear to be present in 1881 to the same extent, with only the far north of Scotland differentiated from the rest of the country – and creating its own tight cluster. The 1881 results show the Shetland Islands and Moray Firth share more in common with the far North of England than with Scotland. The commonality between Southern Wales and the Welsh borders seems to have persisted since 1881, although the pattern at that time is much simpler. Gwynedd and Anglesey remain firmly grouped with central Wales. The extent of the Welsh border region into Wales and England remains largely unchanged.

The withinss highlights an additional change in the likeness between districts that share a region between 1881 and 2001. In 1881, an area corresponding to East Anglia is tightly clustered, suggesting relative isolation from its surroundings, yet the cluster disappears altogether by 2001 with the region becoming grouped with the eastern side of England more generally. The south west cluster becomes enlarged between the years and became significantly more compact, suggesting an increasingly distinctive region compared with the rest of Great Britain. This expansion has not included the southern tip of Cornwall as it appears to have separated from the rest of the south west of the country. The withinss values suggest this is not a dramatic transition between clusters as they were very high for the south west cluster (of which Cornwall was a part) in the 1881 result and became much lower in the 2001 result when Cornwall splits off. This indicates an awkward grouping in 1881, caused by relatively large differences between Cornwall and the rest of the cluster.

K-means clustering of English Lasker Distances in 1881 produces a noisy map, suggesting a much greater degree of diversity between English districts at that time, or quite possibly a greater variation in data quality. Central England is especially muddled, but discernable regions exist for the south west and Cornwall, the south coast, East Anglia and the far north of Scotland. In addition, the withinss values in general for 1881 are many times larger than those in 2001 suggesting that there is

greater difference between the 1881 Registration Districts than the 2001 Local Authority Districts. In the context of previous discussion related to increasing migration in the 21st Century, and the impact of more spatial units on the relative dissimilarities between them, this finding makes intuitive sense.

Whilst 15 clusters represents the most interpretable result (in terms of ease of visualisation) in this context, to facilitate comparison with the Ward's hierarchical clustering below and to demonstrate the challenges of attempting to establish the similarity between each cluster Figure 5-9 maps the *K*-means output for 20 clusters using both the 1881 and 2001 data.

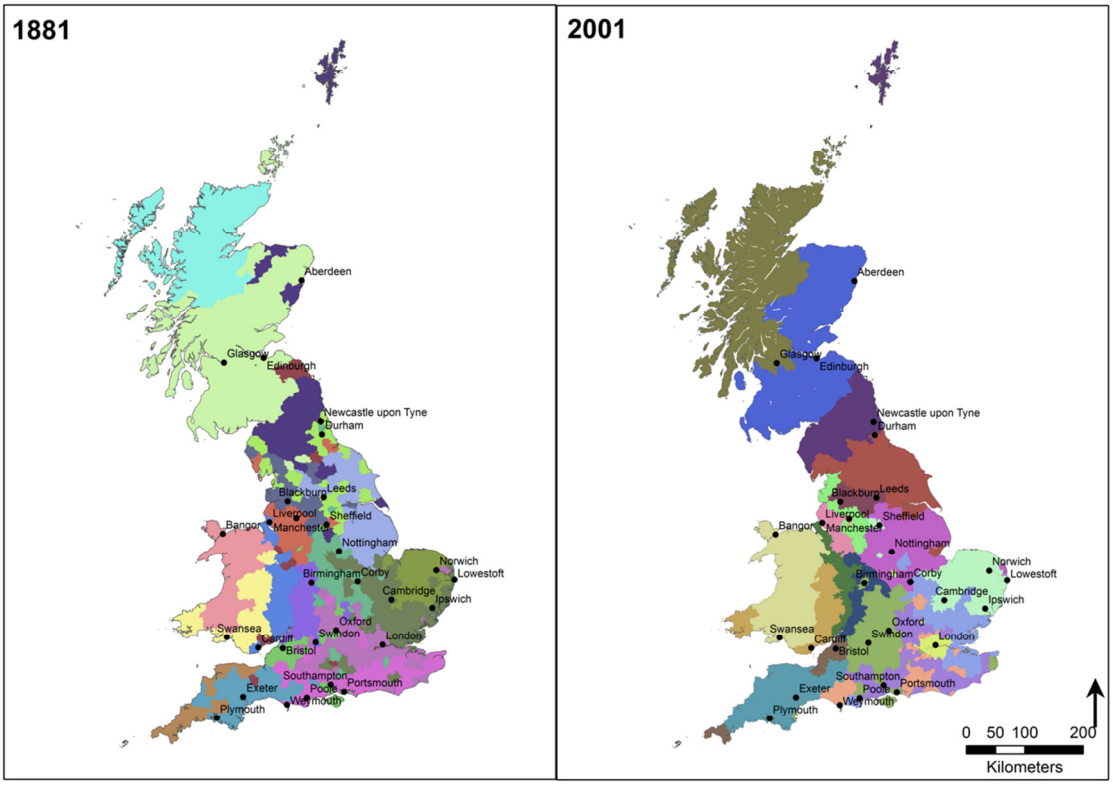


Figure 5-9: Results from *K*-means clustering where *k*=20 for 1881 (left) and 2001 (right). These are produced for comparison with the Ward's Hierarchical Clustering result outlined below. Colours do not correspond between the two years.

5.4.3.3 Ward's hierarchical clustering

The advantage of Ward's hierarchical clustering is that, unlike *K*-means, it creates a nested hierarchy in the classification so that the number of clusters can be amended without changing the global classification result. In addition, the resulting dendrogram provides an indication of the strength of relationship both between and within clusters. For this reason it enables many more aspects of the clustering result to be analysed and greater flexibility in its interpretation. This, combined with the increased coherency of the results, is the reason for its in depth discussion (in comparison to the *K*-means) below.

For the purposes of the analysis reported here (and with the CAS Ward level analysis below in mind) 20 regions were produced because the dendrogram structure became less distinct in subsequent divisions, and because 20 is approximately the maximum number of classes that a user can realistically be expected to interpret in map form (Krygier and Wood 2011).

As Figure 5-10 demonstrates with the 2001 data (see Appendix 1 for the 1881 equivalent) there is a strong spatial patterning to the results. Maps of the resulting cluster outcomes (Figures 5-11 to 5-13) show that Ward's creates compact, homogenous regions from the Lasker Distance data. The first split in the 2001 dendrogram – that is, the largest difference in Lasker Distance between Local Authority Districts – occurs between Scotland and the rest of Great Britain. One exception is the Northamptonshire town of Corby that shares a branch with the Southern Scottish Districts (cluster 3). This is an interesting outcome and one that is discussed in more detail in Section 6.1.1. The next split in the dendrogram produces a single branch (4) that comprises the Welsh districts. As with the Scottish cluster, the demarcation between Wales and England closely follows the contemporary national boundary. The cluster allocation for Wales suggests relatively low surname diversity within the country, as all of its Districts are assigned to a single cluster. The third major split in the dendrogram creates one branch for southern England (clusters 5-10) and a separate branch for northern England (clusters 11-18). The branches (5-7) of south-east England represent three contiguous clusters around

London, and form three of the most populated surname clusters in Britain. Cluster 9 neatly separates Cornwall from the rest of the country, whilst cluster 10 contains the rest of south-west England. Cluster 8 is assigned to the same branch and represents the Welsh border region. The Local Authority Districts classified in cluster 11 are interesting as they fall between Nottingham, Birmingham and Peterborough, but do not include Leicester. This area can be characterized by a scattering of small settlements that nonetheless share similar cluster values.

Cluster 11 appears as a border region in surname composition between northern and southern England and, based on its position in the dendrogram, has slightly more in common with the former. Clusters 12-16 form the north Midlands and the southern extremity of northern England. Clusters 13, 14 and 16 occupy the largest spatial

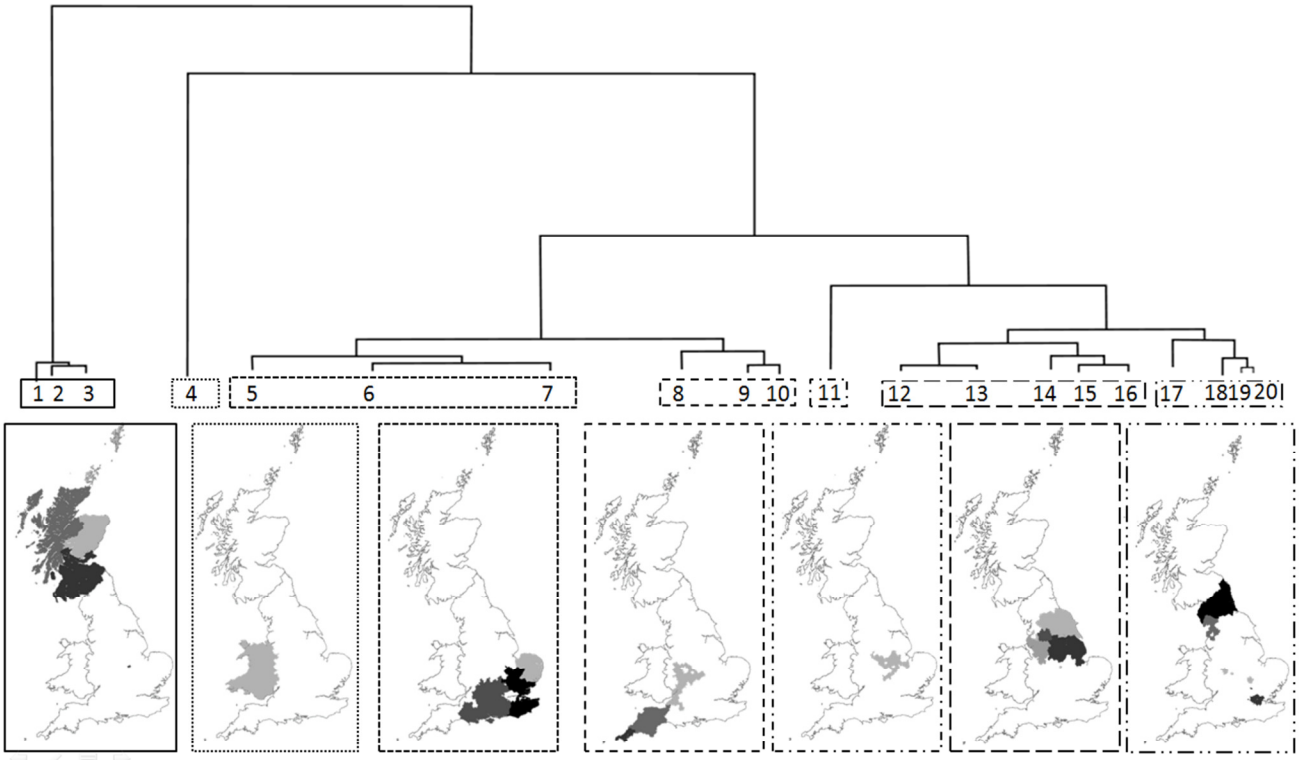


Figure 5-10: The dendrogram produced from Ward's hierarchical clustering of the 2001 Lasker distances calculated at Local Authority District level with seven aggregations of the resulting 20 cluster allocations mapped along the bottom. Each branch of the dendrogram shows a clear spatial pattern: clusters 1,2 and 3 are all Scottish; cluster 4 is entirely Welsh; clusters 5-7 make up southern and south-east England; clusters 8-10 make up western and south-west England; and clusters 11-18 together make up northern England. Clusters 19 and 20 are Birmingham and London respectively. Printed in Longley *et al.* (2011a)

extents: 13 includes the cities of Manchester and Liverpool to the west, while the eastern cluster (16) includes Nottingham and Sheffield and cluster 14 extends furthest north. These cities are not separated from their hinterlands by the clustering process, suggesting that the regions are more distinctive than the high order settlements upon which they are focused. Cluster 12 is one of the tightest spatial clusters produced at this level, and comprises an area of predominantly small, scattered settlements but is surrounded by large towns and cities (Manchester, Leeds, Blackburn). Cluster 15 is relatively small in areal extent and is spatially fragmented. Clusters 17 and 18 make up the remainder of the far north of England, truncating abruptly at the Scottish border. The final two clusters are anomalous in that they represent one city each: Birmingham (19) and London (20). The proximity of these to one another in the dendrogram, coupled with their assignment to separate clusters suggests that their status as the two largest metropolitan areas (as opposed to conurbations) in Great Britain is manifest in distinctive surname characteristics (McElduff *et al.* 2008). To establish a comparative geography of the major cluster divisions for each year (those occurring closest to the top of the tree) multiple maps were produced by increasing the number of dendrogram divisions from two to seven clusters (see Figures 5-11 and 5-12).

When comparing 1881 to 2001 (in Figures 5-11 to 5-12) the first cluster division is one of the most interesting as it suggests that Wales has increased its relative similarity to England, and Scotland has become more different as the first split in 1881 forms between England and Wales, whereas in 2001 this split occurs between Scotland and England. It is not until the fourth split that Scotland is partitioned from the rest of Great Britain, suggesting a greater difference between North and South England in 1881 than Scotland and Northern England. The North/ South split in 2001 occurs at the fourth split and slightly further North of its position than in 1881. The level with five clusters differentiates the far North of England from the combined Northern/ Midland areas in both years, although the partition is located further north in 2001.

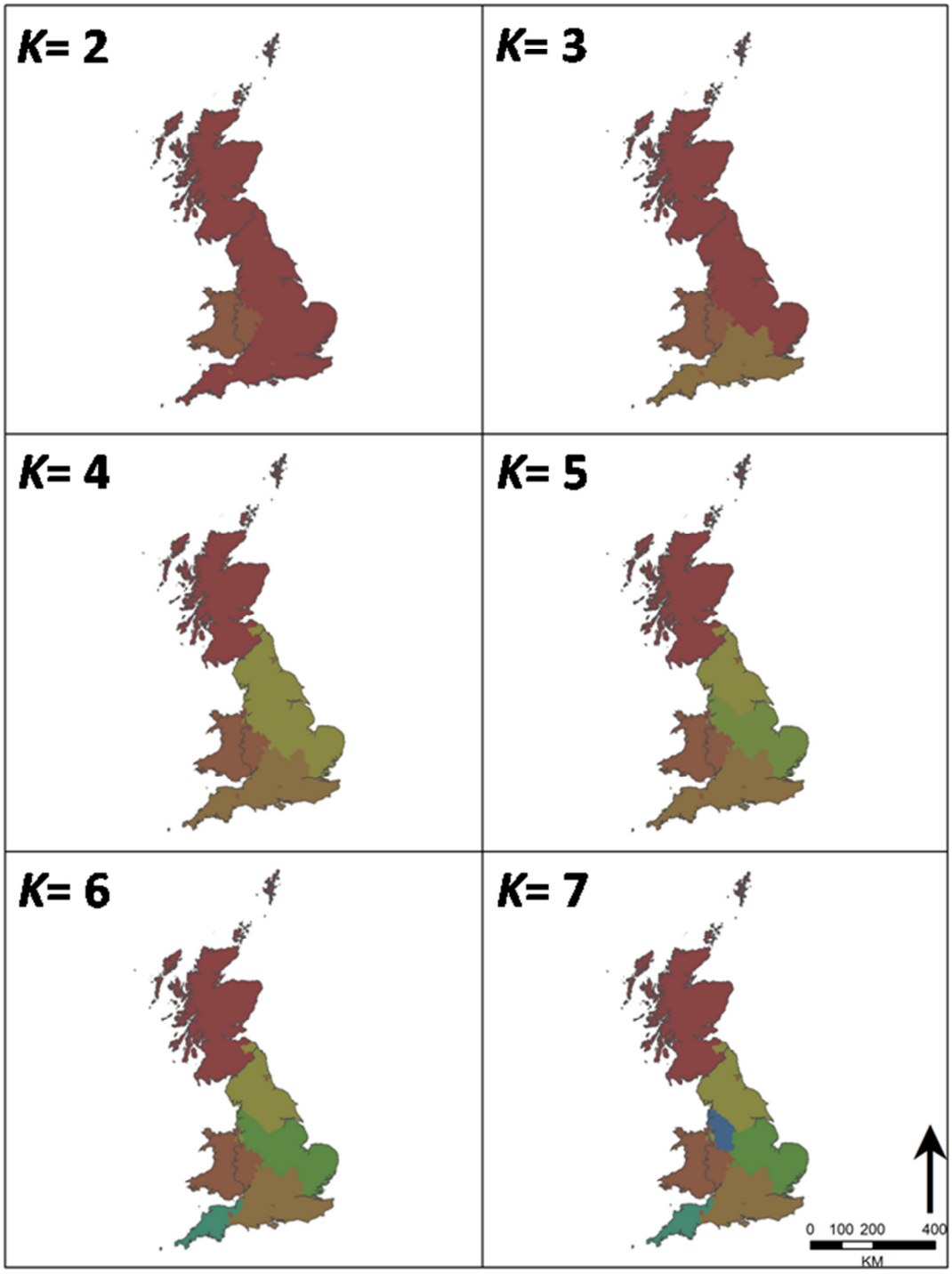


Figure 5-11: Maps of *K*= 2 to *K*= 7 Ward's clusters of the 1881 Lasker Distances. Wales becomes distinctive at *K*= 2 clusters, there is a North/ South split in England before Scotland becomes highlighted at *K*= 4 clusters. Southwest England is distinguishable at *K*= 6 clusters.

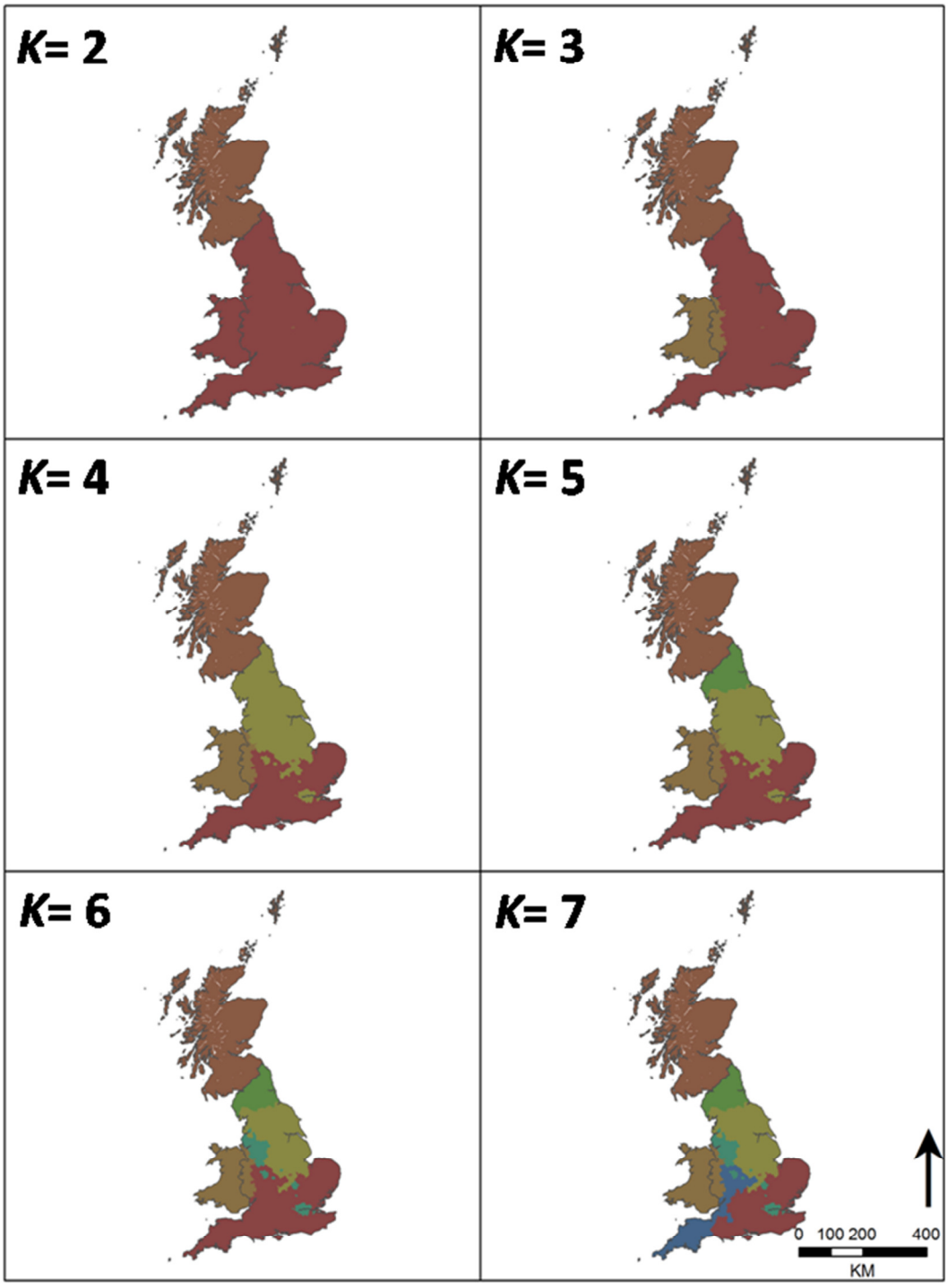


Figure 5-12: Maps of *K*= 2 to *K*= 7 Ward's clusters of the 2001 Lasker Distances (with London as a single district). Scotland becomes distinctive at *K*= 2 clusters, Wales appears at k=3 before a North/ South split in England occurs at *K*= 4 clusters. Southwest England is distinguishable at *K*= 7 clusters.

The cities in the North West and London create the 6th cluster in 2001; a position occupied by the Southwest in 1881. The former, excluding London, appear at the 7th cluster in 1881 and an enlarged Southwest area, including along the Welsh borders and Bristol Channel are distinguishable by the 7th cluster in 2001.

Figure 5-13 overlays the 2001 regions (shown in Figure 5-10) on top of those produced for 1881. It shows that at 20 clusters the surname regions of 1881 and 2001 represent very similar patterns. Notable exceptions include the division of Scotland between the highlands and lowlands (including the Scottish islands), the

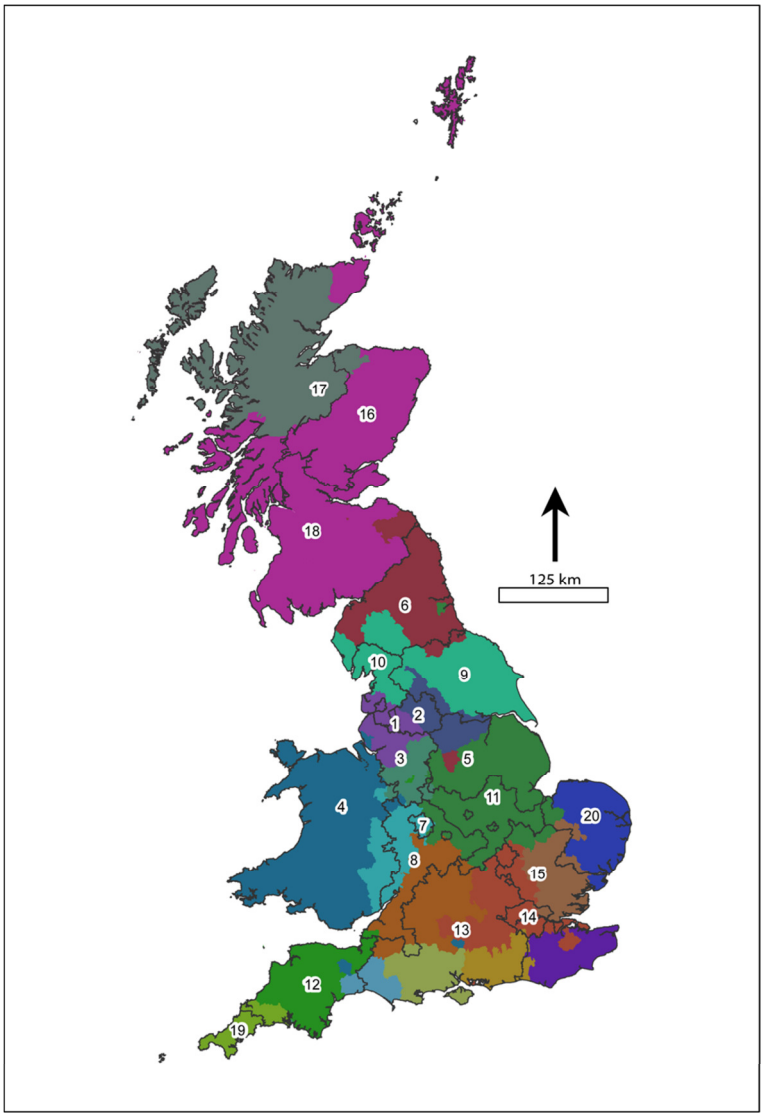


Figure 5-13: A comparison between the 1881 Ward's Hierarchical Clustering result (unique colours) and the 2001 equivalent (solid lines). The latter have been numbered. It is clear that there is a surprisingly close resemblance between the two years.

spread of the Welsh region into England in 2001 and greater differentiation within the South West in 1881. In addition it is clear that a number of urban areas have been isolated in 2001 which have not in 1881, suggesting an increase in diversity (relative to the rest of the population) between the two years.

Of the two clustering methods discussed above, Ward's hierarchical clustering appears the most effective in this context. The regions produced are contiguous and meet expectations based on known transitions within Britain's population structure.

5.4.3.4 Multidimensional Scaling

The analysis described above provides a discretised view of the differences in surname composition across Great Britain. MDS presents a more continuous impression and appears to reinforce the general assertion that areas that are further away from each other geographically are likely to be characterised by increasingly diverse surname compositions.

The maps produced from the MDS coordinates (Figure 5-14) agree broadly with the clustering outputs. 1881 presents a noisier picture with several regions standing out from those contiguous with them. Both maps, especially from the 2001 data, illustrate a gradual change from the north to the south or east to the west of the country, with the most abrupt changes occurring at the present national borders between England and Scotland and England and Wales. Northern England appears more similar to Scotland in 1881 compared to today where it exhibits a strong difference from both Scotland and the rest of England. Based on the colour changes in the 2001 map, one can split Great Britain into the following regions in 2001:

1. Northern Scotland
2. Southern Scotland
3. Far North England
4. North West England
5. Wales and England/Wales border region
6. East Anglia
7. Central England.

8. Cornwall and the South West.

These regions appear much less clear in 1881. At this time, Great Britain could be split into:

1. Scotland and the Far North of England
2. Wales and England/Wales border region
3. South West England and Cornwall
4. North/Central England
5. Southern, Eastern and Central England
6. Many relatively unique districts throughout Great Britain.

The MDS scatter plots shown in Figure 5-15 attempt a more literal representation of the data used to produce the maps described above. The 1881 MDS plot in the XZ dimension (Figure 5-15B) shows how Scotland (red) and Wales (blue) appear at opposite ends of the distribution and appear with few other districts in their point cloud. The ZY plot of the 2001 MDS coordinates (Figure 5-15C) highlights the clustering of districts from the North West, Wales and West Midlands. All plots show that districts closer to each other are likely to have more similar Lasker Distances.

The general rate of change across Britain appears to be broadly similar for both 1881 and 2001. The most obvious exception to this is the Anglo-Scottish border that appears to have become more pronounced over the course of the century. On the basis of the finer scale analysis below this is not thought to be an artefact of the spatial units used. The ability to highlight both gradual and abrupt changes is a key advantage to mapping the MDS values in addition to plotting in the conventional way. The method therefore provides useful context to the always discrete boundaries produced by clustering.

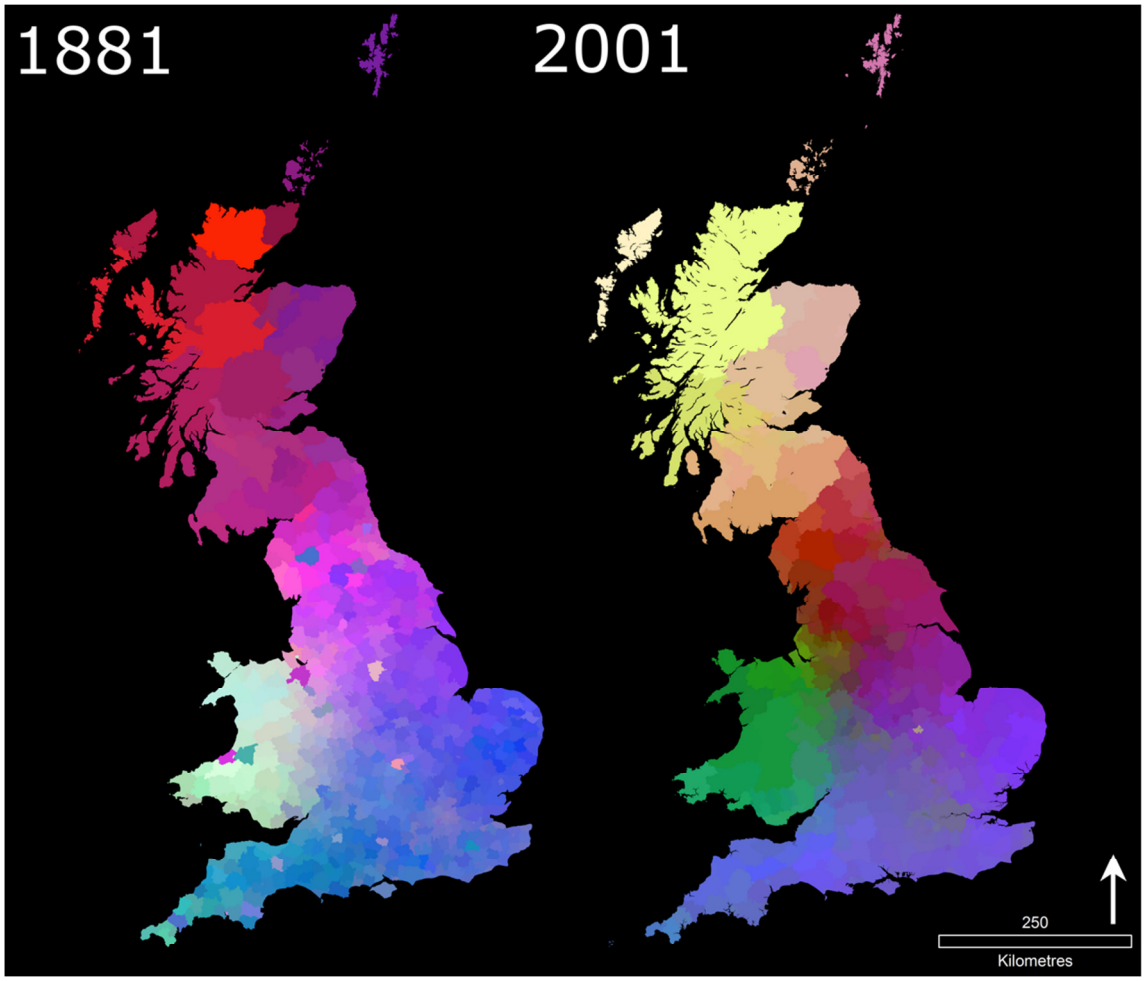


Figure 5-14: MDS Maps demonstrating both the abrupt and gradual transitions in surname composition across Great Britain in 1881 (left) and 2001 (right). Colours are not equivalent between the two years but the magnitude of variation between hue and colour intensity are comparable.

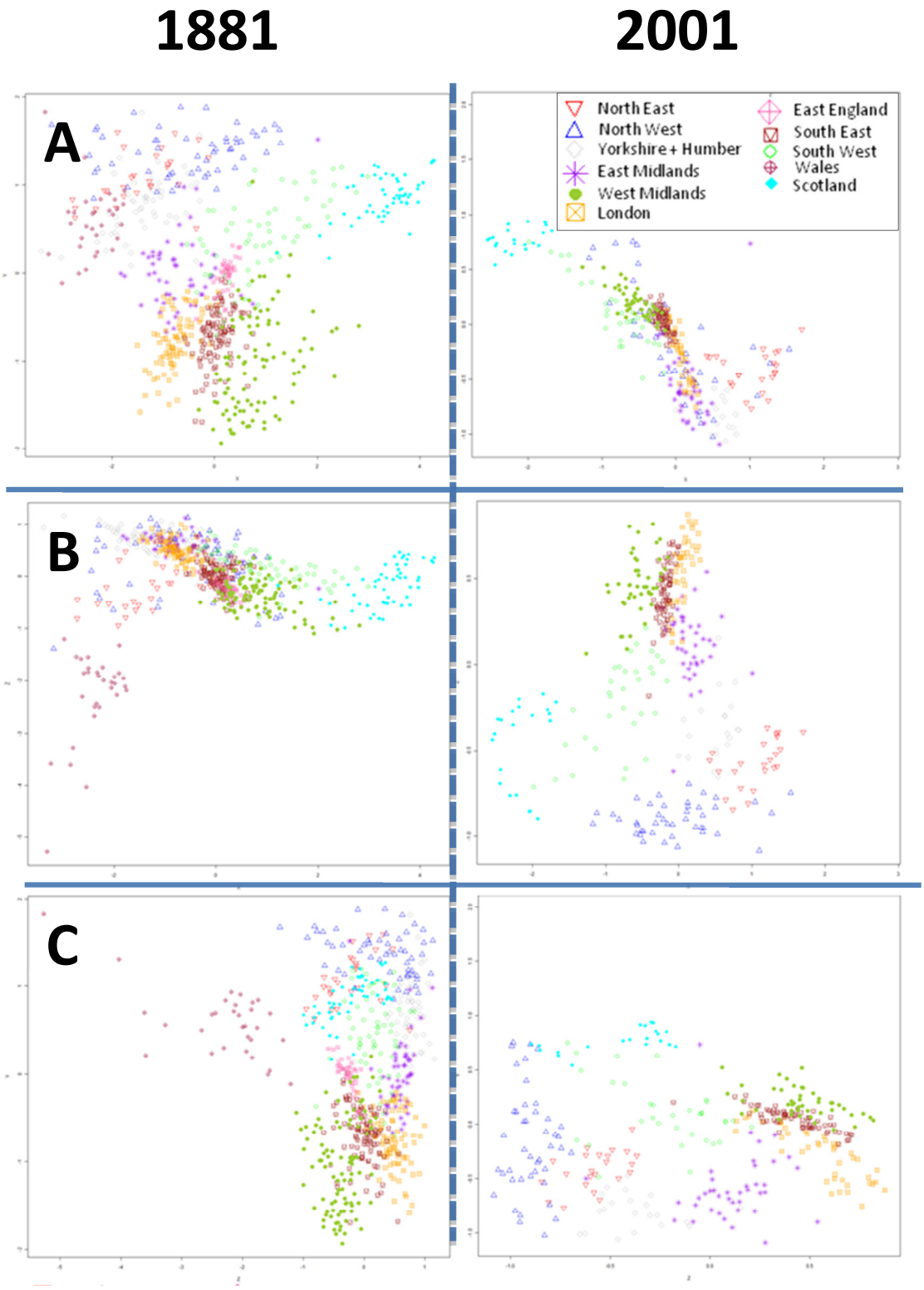


Figure 5-15: MDS results plotted on the YX(A), XZ (B) and YZ (C) axes. The colour and symbol of each point represents the Government Office Region (GOR) that the Local Authority District falls within. The plots demonstrate the clustering of Local Authority Districts that share a GOR. Those closer together will have more similar colours in Figure 5-14.

5.4.4 FINER SCALE ANALYSIS: MDS AND WARD'S HIERARCHICAL CLUSTERING

A central aim of this work is to identify broad regions that share internal similarity in their surname composition. As is demonstrated above, Ward's hierarchical clustering is the method that best achieves this, as it not only clearly identified broad regions, but also adequately captured known anomalies in Britain's surname distribution, such as the Scottish surnames in Corby. The number of clusters represented can be easily varied without having to re-cluster the data. However, while *K*-means clustering can be used to create hierarchies, the exact nature of the hierarchies may differ between each set of repeated runs because of the effects of initial random seeding. In addition, the MDS visualized as a graduated color map provides a powerful visual indication of both the gradual and abrupt trends in surname similarity between areas of Great Britain. It therefore serves as a reminder that not all boundaries are as abrupt as those implied by cluster analysis and complements the discrete regions produced by Ward's. The analysis described in this section has been undertaken on the 10,500 CAS Wards from the 2001 Electoral Roll and implements the two most successful methods outlined above. It represents by far the most detailed analysis of surname distributions ever undertaken.

5.4.4.1 Wards Hierarchical Clustering

As Figure 5-16 shows, partitioning the data into 20 clusters using Ward's hierarchical clustering at the Census Area Statistics (CAS) Ward level produces similar spatial distributions and cluster extents to those derived from clustering the Local Authority District data (see Figure 5-10). The two most noticeable exceptions are the different allocations in Scotland and Wales. The former is partitioned into twice as many groups with separate groupings for central Scotland around Glasgow and the islands of Orkney and Shetland. Wales is partitioned into two regions that separate the north from the south. To compensate for these additional clusters, southern England and East Anglia are comprised of fewer groups. As such, the finer detail provided by the CAS Ward level data serves to differentiate between rural and urban areas.

Using the Local Authority District level classification, London and Birmingham were each assigned a single unique cluster, while at CAS Ward level these clusters of surnames are observed to occur in metropolitan areas throughout England, and in Glasgow (Figure 5-17). This is one of the great strengths of the finer scale

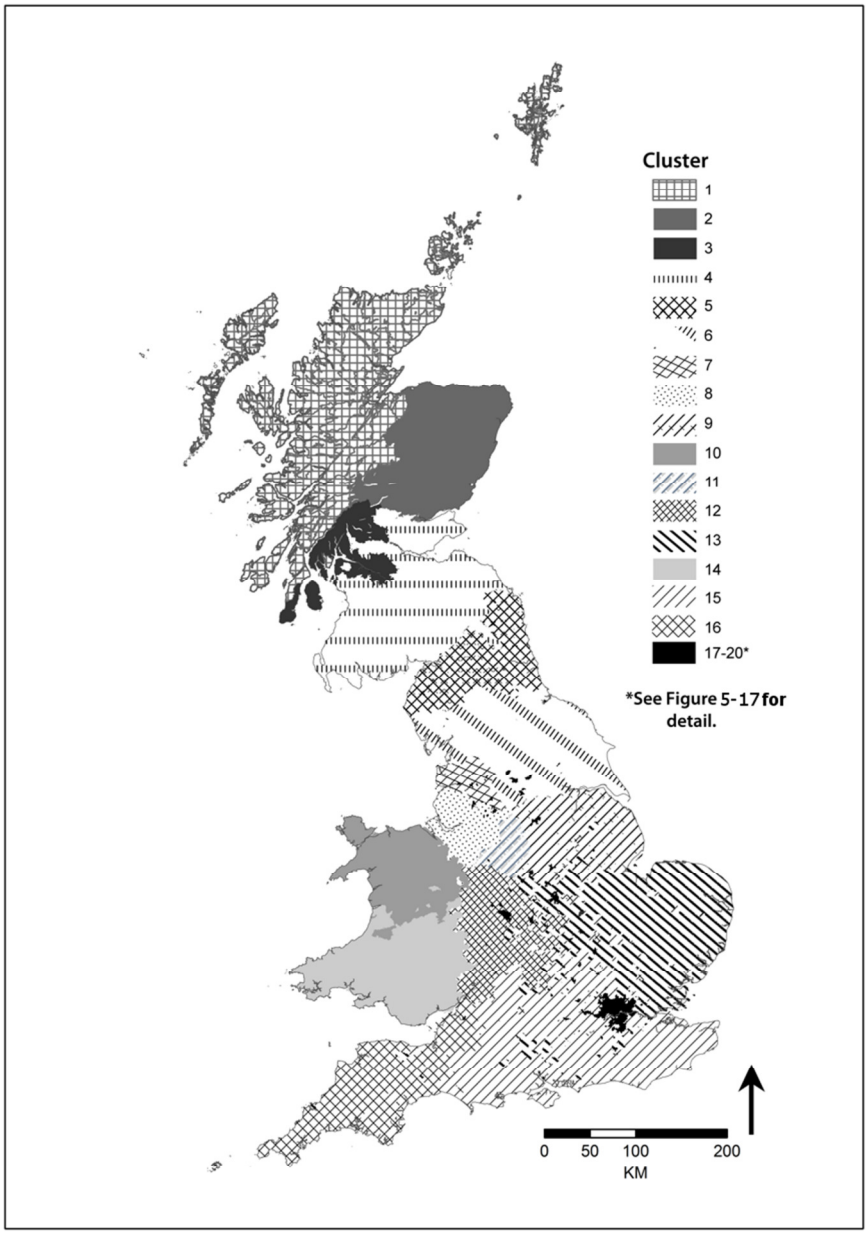


Figure 5-16: Map showing all 20 cluster allocations derived from Lasker Distances, calculated at the CAS Ward level. This shows a close correspondence with the Local Authority District level clustering in Figure 5-10. Published in Longley *et al.* (2011a).

classification produced with smaller spatial units as it begins to disaggregate the impression that metropolitan areas conform to the broad regional characteristics of their hinterlands. A reasonable conclusion from Figure 5-17 is that some areas of Britain's towns and cities share more surnames in common with areas tens (if not hundreds) of kilometres away than those contiguous with them. This is further enforced by Figure 5-18 (produced using www.wordle.net), which shows the distinctive surname characteristics for each of the clusters. The size of the most popular 15 names is scaled according to the frequency of their occurrence and the numbering on this figure corresponds to the regions identified in Figure 5-16, and it is clear that the clusters are very distinctive in terms of their surnames composition. This distinctiveness in all but the "urban" clusters is clearly linked to spatial proximity with areas closer to each other more likely to share similar surnames.

The use of less aggregate data therefore appears to unearth an additional level of population structure visible from the distributions of surnames. Clustering data at CAS Ward level gives the impression of large-scale generalized trends punctuated by smaller groups associated with urban areas. The former appear to be the product on the uniting/ isolating effects of distance whilst the latter reflect social behavior as migrants move to geographically disparate, but culturally similar, urban areas. This conjecture is reinforced by the MDS results below.

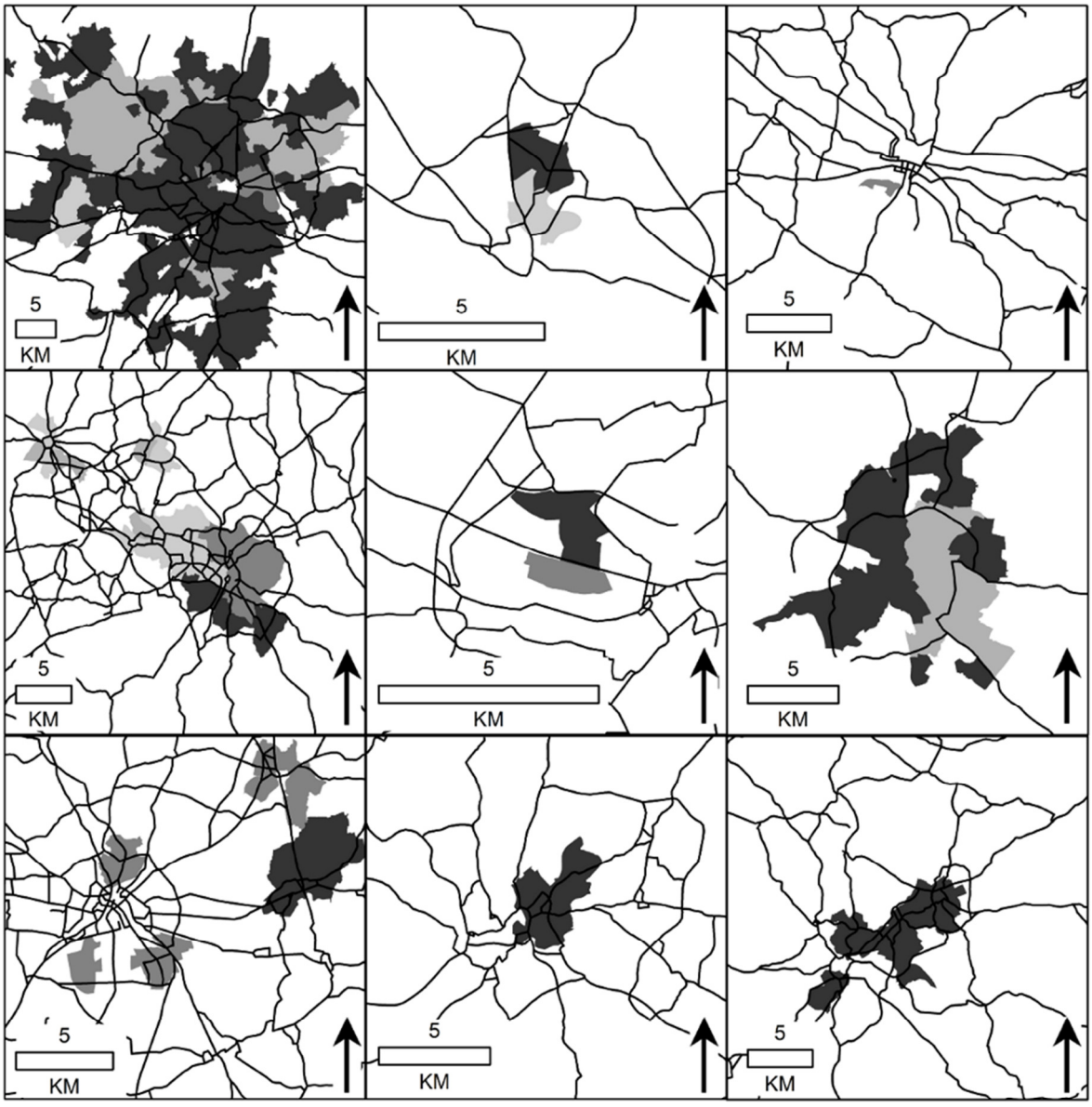


Figure 5-17: Maps illustrating the similarity in surname composition between nine urban areas in Great Britain, produced using Ward's Hierarchical Clustering (20 clusters) of Lasker Distances at the CAS Ward level. The three common cluster allocations are shown by the differing shades of grey, major roads are indicated as black lines. All other cluster allocations are white. From top left to bottom right the areas are as follows. London; Southampton; Glasgow; Birmingham; Newcastle-upon-Tyne; Leicester; Manchester; Bristol; Sheffield. Published in Longley *et al.* (2011a).

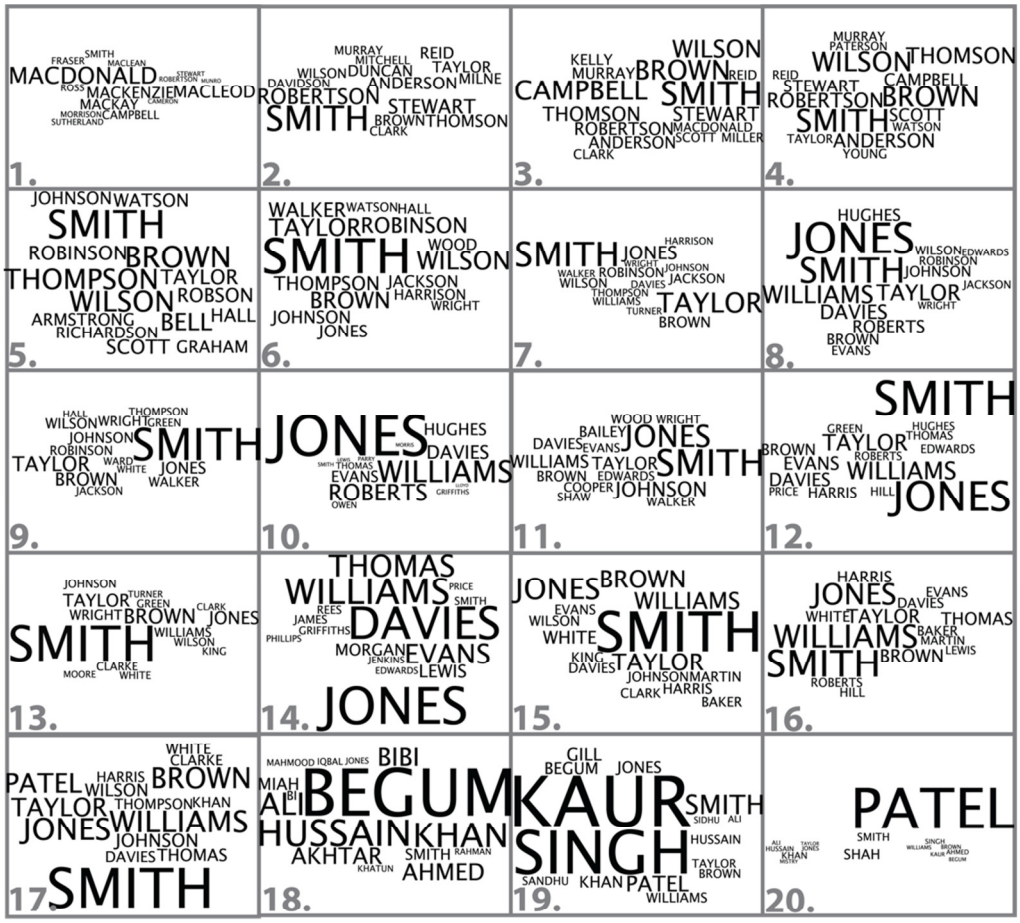


Figure 5-18: Small multiple Wordle, showing the 15 most commonly occurring surnames in each cluster, with size of lettering scaled according to absolute frequency. The clusters are numbered as in Figure 5-16. Published in Longley *et al.* (2011a).

5.4.4.2 MDS

To this end, Figure 5-19 presents two of the three combinations (X vs Y; X vs Z; Y vs Z) of MDS coordinates derived from the CAS Ward level data. By sub-dividing the data into the 9 English Government Office Regions (GOR) plus Scotland and Wales, it is clear that CAS Wards are distinctively clustered in multidimensional space. The plots show that each GOR is tightly distributed and, with the exception of Scotland and Wales, can be characterized by the majority of CAS Wards presenting values towards zero in the *Z* dimension. The distributions are also clustered in the *X-Y* dimensions, with the head and tail of each cloud of points for each GOR showing only very limited overlap with the other GOR point clouds.

Identifying the points closest to 0 in the *Z* dimension shows these to be the key contributors to the distinctive regional characteristics identified by the hierarchical clustering analysis. Regions with large numbers of CAS Wards that differ from the general surname composition of a region have more points closer to the negative end of the spectrum. This is illustrated by the shape of London's point cloud in comparison with those of Wales and Scotland. In the XZ dimension, Scotland has the least concentrated distribution of points, which may be because of the large variation in the surname composition of the Scottish Islands when compared to the mainland. Unsurprisingly given the hierarchical clustering results, Southwest England shows a high concentration of points in all dimensions, suggesting similar surname compositions throughout the CAS Wards.

Analysis at Local Authority District level (see above) produces similar clustering within GORs, but without the long tails that is observed in the distributions of data points at CAS Ward level (see Figure 5-15). This is an artefact of the increased generalisation that arises when using larger spatial units (in this case Local Authority Districts): as suggested above, the surname diversity in the urban areas is only clear at the CAS Ward level of geography and will, therefore, not produce the long tails at Local Authority District level. Although the use of GORs in Figure 5-19 (and Figure 5-15) should not be taken to imply that the GOR is anything more than a practical means of partitioning the c.10,500 data points in the MDS plot, these results do give credence to the view that this prevailing administrative geography is surprisingly consistent with the surname geography of Great Britain.

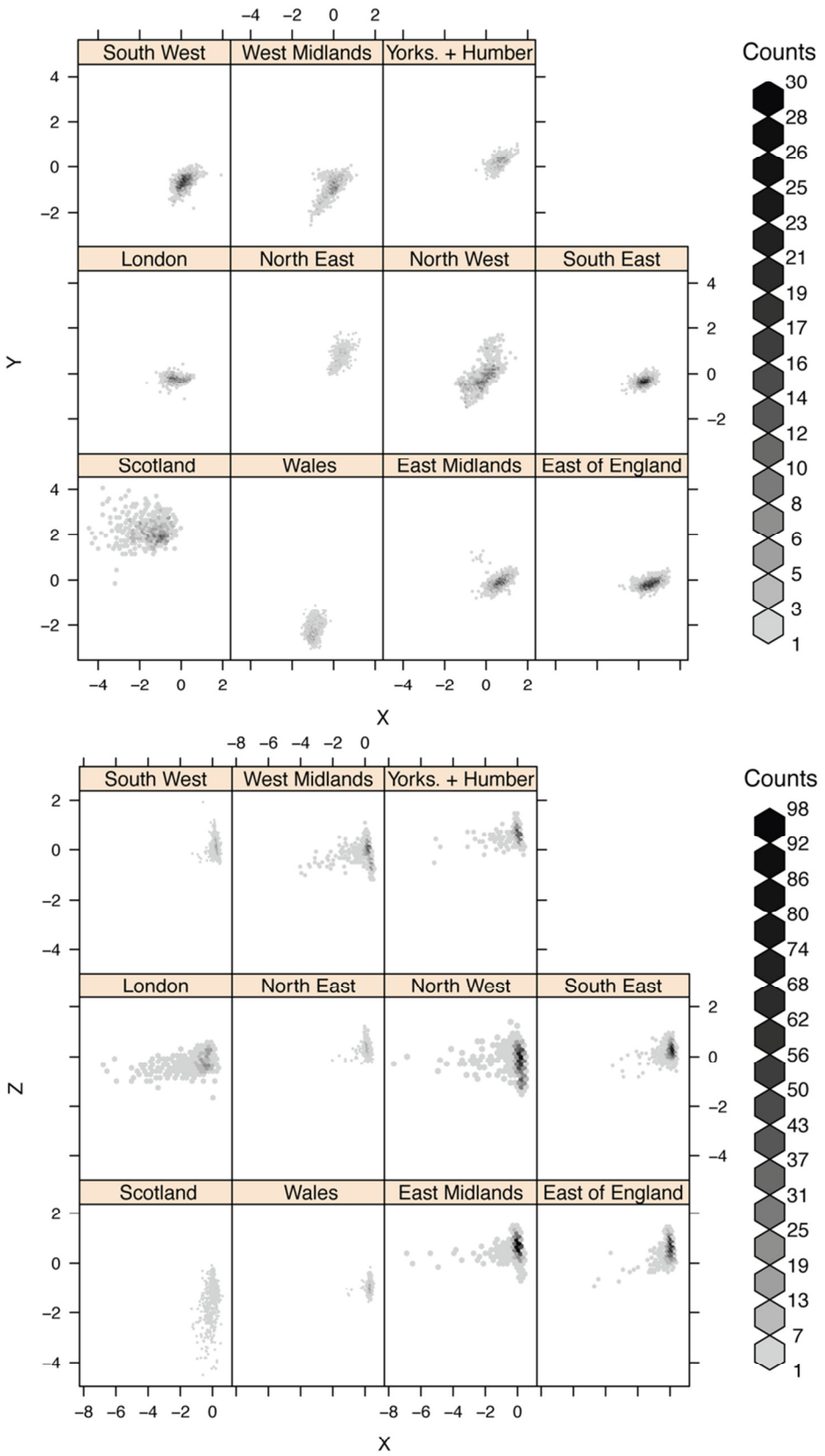


Figure 5-19: Hexagonal binning plots showing the *X-Y* and *Z-X* views of the three dimensions produced by multidimensional scaling of CAS Ward level Lasker Distances. A plot is produced for each Government Office Region (GOR) where each data point represents a CAS Ward. Subsetting into GORs and hexagonal binning were used to ease interpretation of the large number (>10,500) of data points. Each of the plots demonstrates the relatively tight clustering of Lasker distances within each GOR. Published in Longley *et al.* (2011a).

Figure 5-20, perhaps the most detailed map in this thesis, maps the MDS values for the c.10,500 Wards in Great Britain. The effect of distance is clear with a gradual transition from blues and greens in the south to reds further north. The Scottish border is clearly defined as an abrupt transition from greens to reds, whilst the Welsh border is more diffuse by comparison. Urban areas are picked out in browns, the Scottish Islands appear greener with distance from the mainland and there is a more subtle but abrupt shift from blues to greens between the Southeast and Southwest of England. A possible explanation for this latter observation is provided in Section 6.1.2.

The effectiveness of the map in Figure 5-20 lies in its level of detail. It offers compelling evidence for two aspects to modern surname distributions in Great Britain. The first is the underlying trends produced by the surnames that remain concentrated near to, or within, their places of origin in England, Scotland and Wales. These comprise the "base" level surname distributions responsible for the broad regions identified for both 1881 and 2001. Overlain on these are the impacts of migrations from both within and outside Britain that are seen by the brown colours corresponding to urban areas. The method has even been able to capture the between and within urban similarities (such as a split between East and West London) shown by the clustering in Figure 5-17. Mapping MDS values is limited when trying to quantify, non-visually, differences between areas, but the original Lasker Distance matrix can be revisited for this purpose or the data can be clustered using Ward's hierarchical clustering or similar. In summary, the map in Figure 5-20, and its associated plots in Figure 5-19, demonstrates that clustering, although useful, can provide only a partial picture of the dual processes of movement and stability in surname distributions and should be used in conjunction with a more continuous representation, such as MDS.

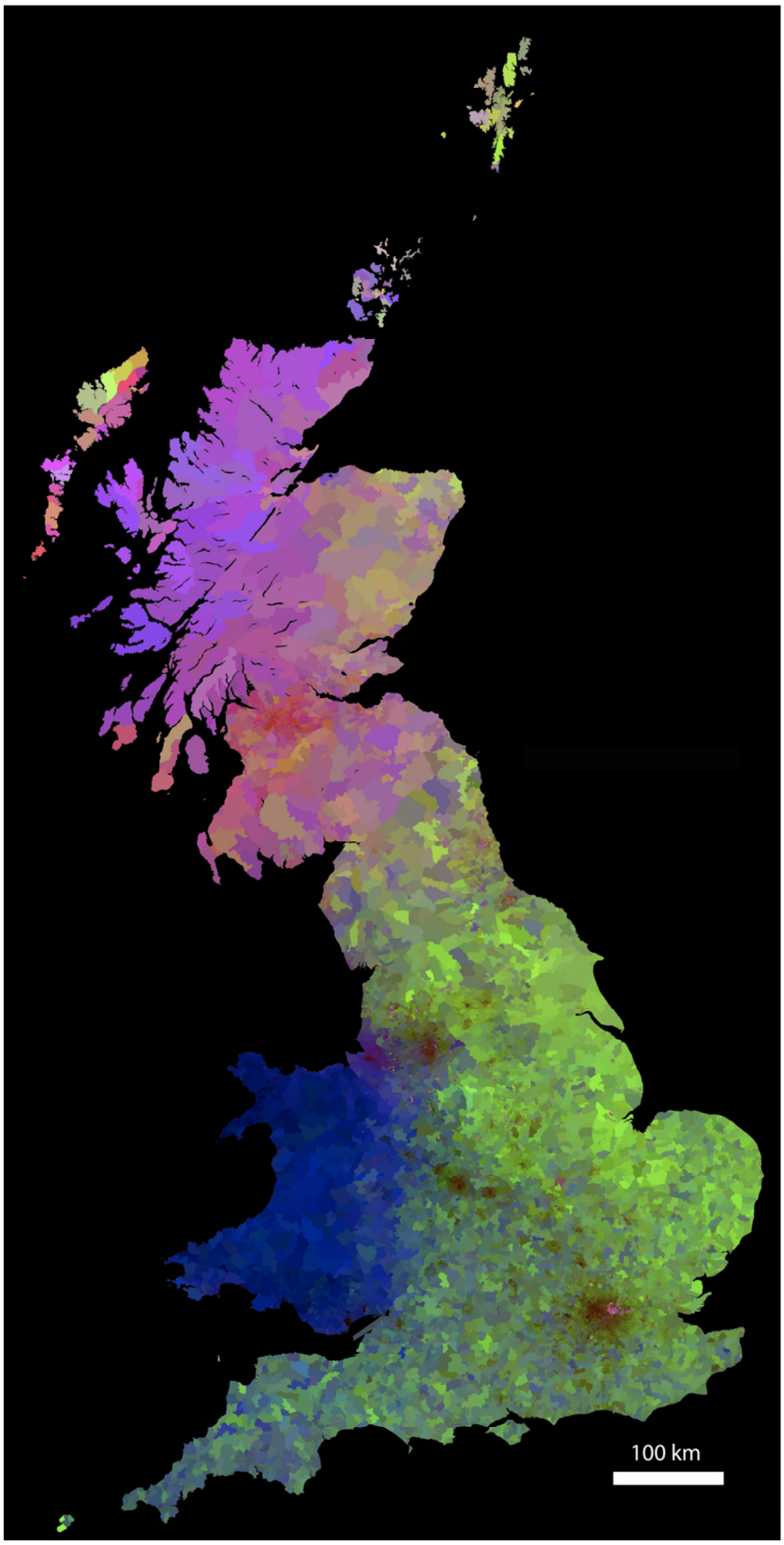


Figure 5-20: MDS map produced with the CAS Ward level data. It shows in unprecedented detail the relationship between geography and surname composition in Great Britain.

5.4.5 DISCUSSION

This chapter has introduced a range of methods appropriate for the quantitative analysis and discovery of surname regions in Great Britain. In addition, the comprehensiveness of the datasets, combined with a temporal dimension and the breadth of techniques used make this chapter the most detailed quantitative analysis of surname regions undertaken at a national level. The following section provides further insights into the analysis above and serves to gauge the effectiveness of the approach to the study of surnames.

The results presented here concur to varying degrees with the limited published research on surnames in Great Britain. Lasker and Mascie-Taylor (1985) demonstrated a north/ south and east/west variation in the frequencies of a selection of surnames found in Great Britain. This chapter has sought to aggregate these unique variations for every surname occurrence in Britain. It has served to validate Lasker and Mascie-Taylor's (1985) paper by showing that such regional trends (as opposed to simple compass direction biases) exist for the majority of surnames and, through clustering and MDS, has illustrated that many surnames continue to share similar spatial distributions. Many of these distributions were highlighted by Guppy (1890) whose 19th Century surname regions provide a surprisingly close match to the findings outlined above (see Section 6.1.2 for in depth discussion). Sokal *et al.* (1992) undertook a very different, but more contemporary, study with much less comprehensive data. In it they suggest a different distribution of surname regions to those produced here, but share the strong indication of isolation by distance between surname populations.

The notable exceptions to the patterns identified in previous research emerge when using the CAS Wards spatial units. Previous studies have not investigated surname patterns at such a fine level of granularity and have therefore been insensitive to the urban clusters illustrated in Figure 5-17. In these instances, additional factors, such as international migration, generate large differences in surname composition between adjacent areas, but greater similarity between the areas affected by similar processes.

These influences have been overlooked by previous research, and provide some of the most interesting findings.

The regions created are the outcome of inductive generalisation on the geography of surnames premised on the idea that most of the individuals that share a surname do not move. The general message of the fact of the regionalisation is that characteristics of people and places in Great Britain remain tightly intertwined, suggesting permanence and continuity of regional identity. The similar regions produced from both the 1881 and 2001 data support this.

The larger Local Authority District-level spatial units in 2001 appear to serve as a filter, removing much of the local level variation and producing a generalised impression of British surname geography (as demonstrated by the effect of amalgamating London's Districts). It is therefore reasonable to assume that the map of Great Britain's regions remains dominated by surnames of Anglo-Saxon origin that were first coined over 700 years ago. Analysis of finer spatial resolution data (CAS Ward level) provides additional detail by detecting smaller areas of Britain with surname compositions that have been affected by more contemporary (20th Century, and recent) migration. These do not dominate any part of the over-all map (Figure 5-16), however, because they comprise densely settled urban areas: they typically occur within and around cities and suggest that, in terms of surname composition, British cities may have much more in common with each other than with their rural hinterlands in the broader regions in which they locate (as shown in Figure 5-17). This level of analysis is not possible with the larger Local Authority District-level spatial units.

In summary, the outcome from this chapter, in common with the theme of this thesis, is that surnames are not spatially random phenomena. Instead they provide important insights into the spatial structure associated with populations. This gives valuable evidence that a large proportion of the British population have remained settled in the areas where their family names were first coined sometime between the 12th and 14th centuries. There are many caveats to this proposition – London is a special case, local settlement geography is more variable in some areas of the country

than others, and the granularity of local naming conventions is coarser in some areas (e.g. Wales) than others. Such caveats are readily detectable in this analysis further adding to the detail provided by surnames in the context of population structure. In addition to extending the classification approach to the European level, the next chapter explores population structure, both contemporary and historical, further through a number of examples pertaining to Great Britain.

6 APPLICATIONS AND EXTENSIONS OF SURNAME REGIONS

Drawing on the analysis from the previous chapter, the focus is both its application and extension. The first section will demonstrate the utility of surname regions in unearthing past migration before outlining ways in which contemporary analysis can be compared to both historical boundaries and historical research. The second, more substantive, aspect of this chapter seeks to provide both a methodological and geographic extension to the analysis of Great Britain's surname regions outlined previously in two ways: first a detailed investigation of the geographical structure of surnames at a continental level in 16 European countries (using data described in Section 3.1.3); and, second, the proposal of a clustering technique appropriate for the pan-European scale. The result is a regionalisation of Europe based purely on the geography of surname frequencies that is key to the search for Europe's cultural regions. The chapter concludes with a general discussion concerning the merits of creating a regional geography based on surnames.

6.1 APPLICATIONS

The inductive approach outlined in the previous chapter has not sought out transitions based on pre-conceived perceptions of different naming conventions. Instead the results can be viewed as a basis for hypothesis generation or to shed further light on known population distributions. What follows are some brief case studies to provide additional insights into a number of the more interesting aspects of the patterns outlined above. The first, pertaining to the town of Corby, provides reassurance that the methods deployed here are sufficiently sensitive to capture documented anomalies in population characteristics. The second and third case studies demonstrate the potential use of the surname regions, especially created from the 1881 Census, in historical research.

6.1.1 CORBY: A SCOTTISH TOWN IN ENGLAND?

One straightforward application for comparisons of the 2001 Electoral Roll with the 1881 Census is to highlight areas that may have been especially affected by migration during the past century. As has been shown in Figures 5-17 and 5-18 urban areas provide an obvious example as they have become increasingly international. Some subtler examples exist of domestic migration within Britain and town of Corby in Northamptonshire presents one such illustration. When mapping the Ward's hierarchical clustering result for K= 2 (Figures 5-11 and 5-12) it is evident that Corby is clustered with the Scottish districts in 2001, but not 1881. The town is also highlighted with the 2001 Monmonier's Barrier Algorithm (MBA) (see Figure 5-6) and multidimensional scaling (MDS) maps (see Figures 5-14 and 5-20). One could infer that a migration event from Scotland has occurred since 1881 to produce a surname composition so similar to that of Scotland. This proves to be the case. In 1932 a company called Stewarts and Lloyds announced a project for a new iron and steel works in Corby. The development transformed Corby from a village of 1,500 people to a new town of 34,000 with 10,000 employed at the works (Pocock 1960). Labour was sourced from the contracting or closing Scottish steel works; where workers had the choice of redundancy or moving south to Corby (Grieco 1985).

Recruitment continued into the 1970s, with Scottish migrants accounting for up to 50% of the incoming population, and up to 57% of inhabitants reporting Scottish origin in some areas (Grieco 1985). Grieco (1985) reports the maintenance of strong links between Corby and Scotland with an annual Highland Games and 55% of all visitors registered in the 1981 Census reporting their origins as Scottish.

The steel industry collapsed in the 1980s; the departure of British Steel left the town "with a severely imbalanced social composition, a labour force with skills inappropriate to the economic activity of the surrounding area [and] poorly placed to attract employers into the area" (Grieco 1985: 16). With such bleak prospects and strong links to Scotland, it is surprising that significant out-migration of the Scottish community in the past two decades has not occurred. That this is not the case presents interesting research questions. For example, how many of the present-day inhabitants of Corby were actually born in Scotland? Here, surname geography also shows its value to identify second and subsequent generations of migrant descendants. Having ancestors that were economic migrants in difficult times can have a lasting effect through generations. This can be disclosed by surname geography, as already demonstrated by Longley *et al.* (2007) in the study of Cornish miners to Middlesbrough and the socioeconomic characteristics of their descendants.

The surname analysis outlined in the previous chapter is clearly not the first to unearth the Scottish link to Corby but it could provide the basis for research into lesser known migrations or cultural links. In addition, the Corby example provides reassurance that the methods used to discern surname regions are sufficiently sensitive so as to highlight relatively small-scale patterns within the broad distributions.

6.1.2 HISTORICAL COMPARISONS

The use of the 1881 Census in this analysis facilitates comparisons with known historical boundaries in addition to historical interpretations and analysis of surname regions. An example of each are provided below in the form of the frequent

occurrence of surname transitions along the Danelaw line and also comparisons with the surname regions posited by early authors such as Henry Guppy (1890).

6.1.2.1 Similarities with Historical Boundaries: the Danelaw line

Aside from those distinguishing Wales and Scotland from England one of the most consistent transitions in surname compositions creates a northeast/ southwest split in England. As Figure 6-1 demonstrates this split is coincident with the southern extent of Viking/ Danish rule, marked by the Danelaw Line, in the 9th and 10th Centuries (Darby 1973). The line follows a well-known transition in place naming characteristics as can be seen in the map of Figure 6-2. The presence of Danish place names, as identified through endings such as "thorpe" and "wick" (see Appendix 2 for full list) ends along the line that runs between London and the River Mersey (near Liverpool). The transitions in surnames are less abrupt but the frequent appearance of a coincident transition along the Danelaw line presents compelling evidence of a relationship. Unsurprisingly, the correspondence with the classification is most evident in 1881, but remains apparent in 2001 at even the finest scales (see Figure 6-3).

Whilst it is unwise to "read too much between the dots" (Keynes 1997) to infer the population characteristics of the area from placenames (in relation to Celtic/ Saxon/ Viking origins) they do provide useful context. To the north of this line it is likely that there was some integration of place naming practices with surnames between the Danish and native populations within Danelaw. Evidence suggests that the spread of Danish names south of Danelaw only took place through the land owning elite and would therefore have had a minor influence on the broader population (Keynes 1997).

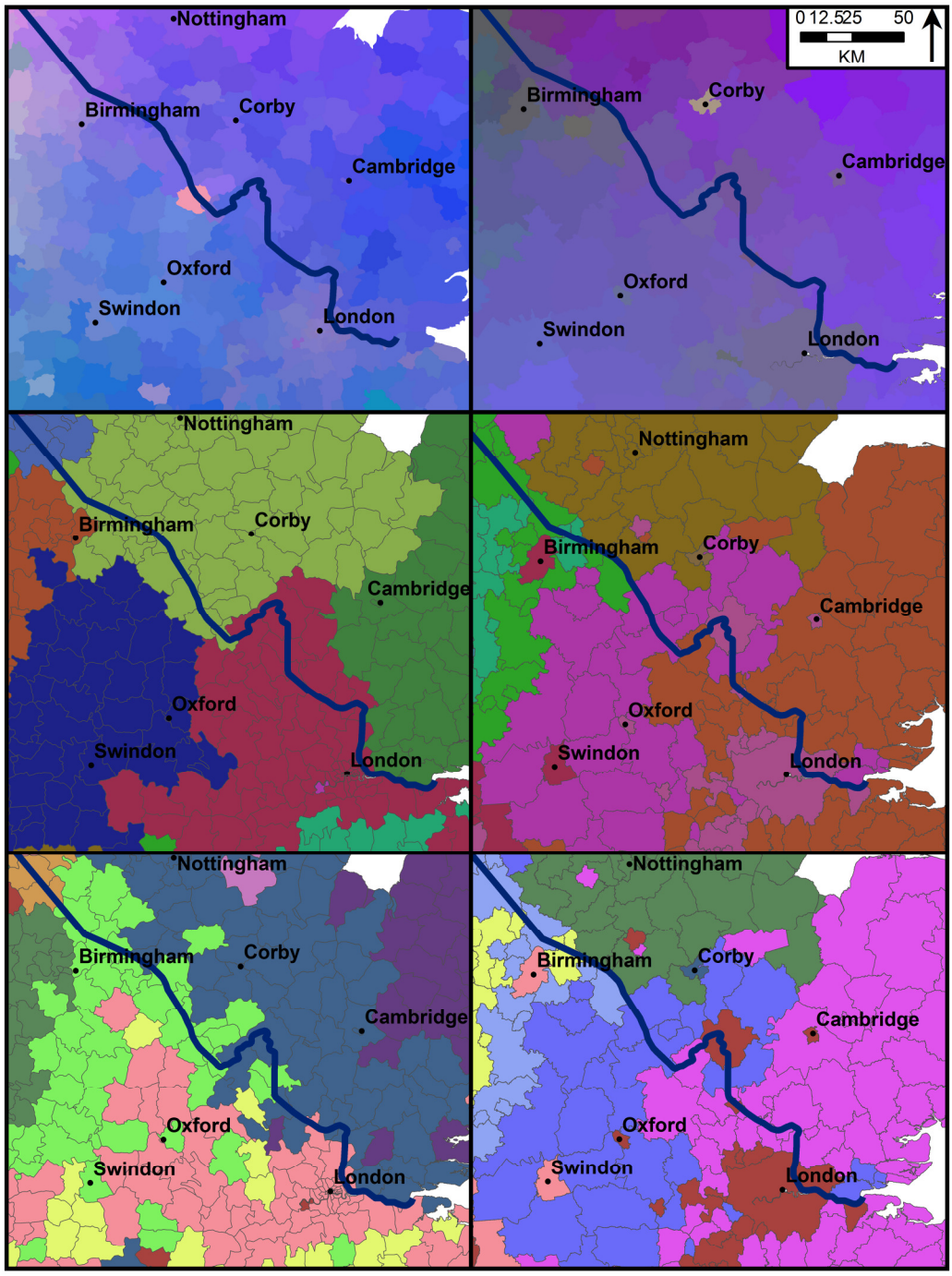


Figure 6-1: Maps demonstrating the correspondence between the path of the Danelaw line and boundaries between surname regions in 1881 (left) and 2001 (right) as identified by (A) MDS, (B) Ward's Hierarchical Clustering and (C) *K*-means clustering.

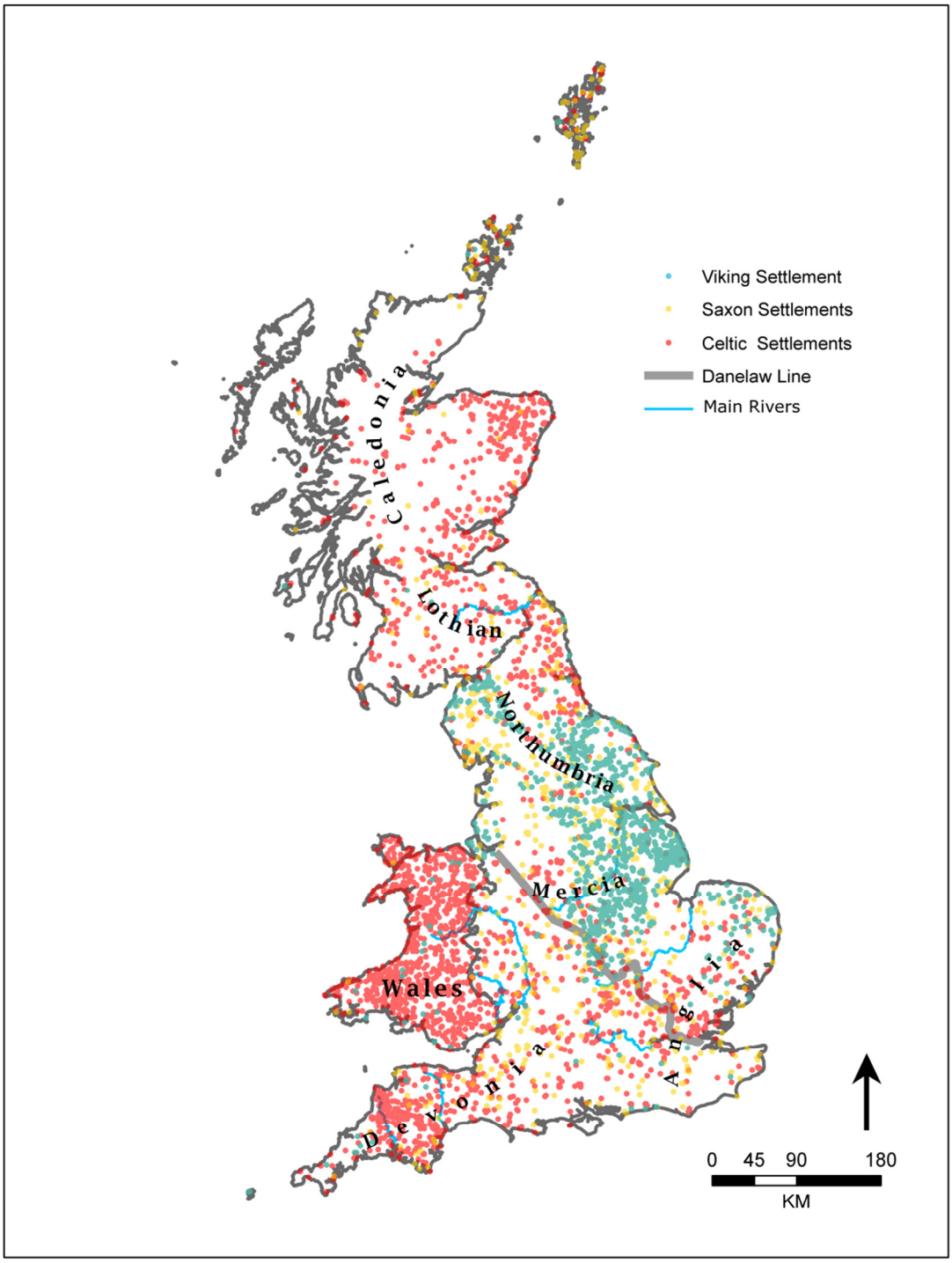


Figure 6-2: The distribution of Viking (blue), Saxon (yellow) and Celtic (red) place naming conventions (see Appendix 2). The Danelaw line is shown in grey and corresponds with the southern extent of Viking naming conventions.

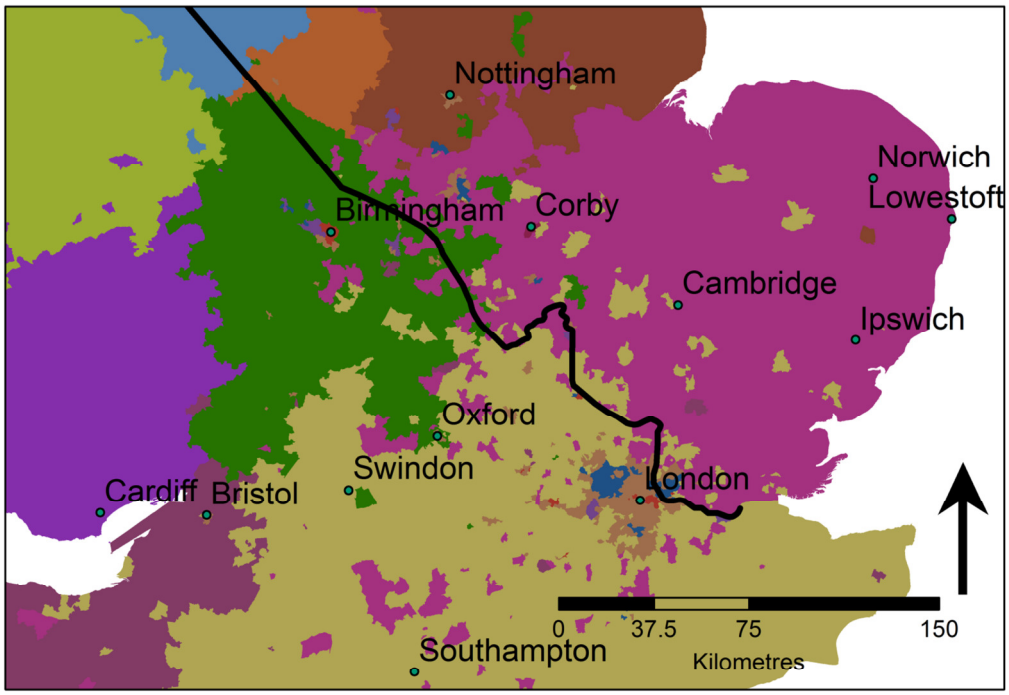


Figure 6-3: The alignment of the Danelaw line (in black) with cluster boundaries produced with Ward's Hierarchical Clustering on the CAS Wards geography for 2001.

It is acknowledged that the link between surnames and Danelaw, whilst visually compelling, requires further research to be adequately proven. It does, however, provide a good example of the sorts of hypothesises that can be generated from the detailed analysis presented here.

6.1.2.2 Comparisons with Historical Research

To further extend the results, comparisons can be made with historical attempts at identifying Great Britain's surname regions. Non-existent computational and limited data resources would in the past have prevented the quantitative approaches used here and made surname research a more manual, deductive and subjective process. Such approaches are extremely time consuming, not consistently applied to multiple datasets and therefore not scalable. Their key strength, however, is the accumulation of contextual information associated with each surname or group of surnames. This information is complementary and validatory to the results discussed in the previous chapter and applies beyond contemporary research. The comparisons below are between the surnames produced by the 1881 Census data and those suggested by authors of the period, specifically Henry Guppy (1890). On this basis it is possible to

compare the 19th Century interpretation with analysis produced using 21st Century methods.

In the opening pages of his book *Homes of Family Names in Britain,* Guppy (1890) outlines his impression of surname regions. Based on his description, these have been approximately drawn and overlaid on the 1881 surname regions in Figure 6-4. Guppy (1890) derived these regions from his extensive work studying the frequency distributions the surnames of yeoman (farmers who owned their land). This group were selected on account of their "stationary habits and purity of extraction" (Guppy 1890: 2). He manually recorded the relative frequencies, taken from the "Kelly's Post Office Directories", by using different sized buttons on a map of Britain (producing maps similar to Figure 4-14). Based on these maps he then provided a largely descriptive account of surname distributions and created the categories of surnames shown in Table 2-3. Guppy's (1890) investigation remains one of the most thorough undertaken.

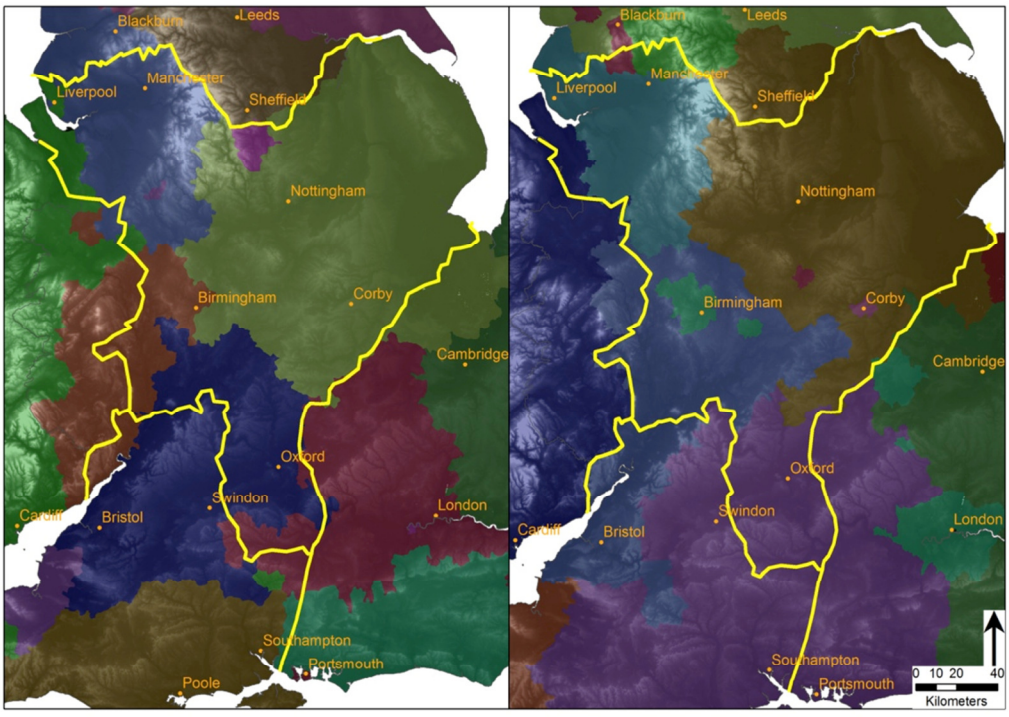


Figure 6-4: A demonstration of the correspondence between Guppy's suggested boundaries for Central England and those created from Ward's Hierarchical Clustering (*K*=15) for 1881 (left) and 2001 (right). SRTM data provides the underlay.

Despite very different methodologies there appears to be a close resemblance between Guppy's (1890) regions and the Ward's hierarchical clustering (15 clusters) of the 1881 Lasker Distances, with transitions occurring along the majority of Guppy's borders. By using a full population register, the results above suggest that "stationary habits" were common throughout the population and not just yeoman as Guppy believed. The agreement between Guppy's interpretation of British surname regions and those of this study is reassuring and demonstrates the utility of inductive generalisation in this context.

It is also interesting to note that many of the themes discussed in this thesis relating to the effects of mobility on the stability of large-scale population structure were also present in the 19th Century. For example, Lower (1860) was concerned that the "locomotive character of the present age" was doing much to "fuse all provincial peculiarities and distinctions" (xxvii). Both he and Guppy (1890) encourage "competent observers in various parts of the kingdom to record the *habitats* [original emphasis] of particular names ere the opportunity now existing shall have passed away" (Lower 1860: xxvii). Fortunately, as this thesis demonstrates, the opportunity has not passed away and it has been possible to create such "habitats" at both the individual and aggregate levels. In addition, and perhaps to the probable surprise of Lower (1860), it is evident that surname regions continue to exist, at all scales, and may only have been marginally weakened by a century of population movements. An important aspect of this research, and one that both Lower (1860) and Guppy (1890) would both approve, is the establishment of a "baseline" regionalisation that has been validated by historical sources and can be used to establish the impacts of change over time. The previous case study, centred on the town of Corby provides a good example of this.

6.2 EUROPEAN EXTENSION

The previous chapter demonstrated how a combination of well-known methods can be applied to create meaningful regions based on surname composition. Confidence in the results is significantly increased by use of high quality data. In the case of the European data, as will become clear, the data are less complete and this must be reflected in the classification methods used. The next section extends the previous approaches by implementing consensus clustering in addition to MDS to produce surname regions on the pan-European scale. Whilst many of the outcomes reflect religious, ethno-cultural or social groups the focus here is simply the discovery of population structure, evident in surnames, and how it is manifest across Europe.

6.2.1 NOTE ON DATA

The surname data, and its associated geography, used in this chapter are outlined in detail in Chapter 3.1.3. A key distinction between the European level data and those used in the study of Great Britain relates to its completeness. The enhanced Electoral Register for Great Britain represents the most complete picture of British population that is readily available. For this reason the resulting regions are likely to be a robust reflection of the target population (that is all residents of Great Britain). In addition a range of geographical units can be tested to establish their impact on the final result. In the context of the European data an element of uncertainty has also been introduced by the different provenance of the surname frequency data for each country. The ultimate data sources for most of the countries are national telephone directories, which obviously do not present identical characteristics across the 16 countries. Such inconsistences may relate to national variations in the gender bias towards male registration in telephone directories, variable penetration of land line rental in the population, different conventions for subscribers removing their entries from directories, different customs in registering names to telephone lines and different procedures and conventions by the companies that commercialised the data. For these reasons, combined with the complications of inconsistent geographic units for each country, the likely regionalisation is going to be less stable.

In addition to this instability, *a priori* knowledge of likely regions, or a reasonable number of groups within the data, cannot be relied upon with a dataset that covers 16 countries in Europe. To decide the optimal number of clusters in this context a greater range of metrics are required than is available in the standard implementations of the clustering routines outlined in the previous chapter. This, combined with the relative (in comparison to Great Britain's regions) instability of the cluster outcome, is responsible for increased uncertainty in the result. To account for this an extension to the standard clustering approaches is suggested below.

6.2.2 DEALING WITH UNCERTAINTY: CONSENSUS CLUSTERING

Indicating the certainty of a clustering outcome is an important aspect of population geography research, especially in regionalisation. There have been a number of previous attempts to do this. For example, Nerbonne *et al.* (2008) used the aggregate data matrices produced in dialectometrics as a basis for identifying linguistic regions. The certainty of such regions were determined through bootstrapping and composite clustering techniques and visualised both as a dendrogram and composite cluster maps. In the former, each branch has information about the number of times a particular grouping between its sub-branches occurred, whilst in the latter lines between geographic regions were drawn with increasingly dark shading, corresponding to the number of times contiguous spatial units on both sides of the line were assigned to different clusters. The mapped results, similar in appearance to Monmonier's Barrier Algorithm (MBA), do not require the assignment of all spatial units to a particular cluster, but the objective is to identify only the most abrupt boundaries.

In this chapter consensus clustering (Monti *et al.* 2003) was chosen as a promising method of creating a robust cluster outcome, consistent with providing a number of metrics to indicate the optimal number of clusters and the certainty associated with each cluster assignment. Such metrics are useful in this context because they provide context to the final clustering outcome. In particular they address the issue only

touched upon in the previous chapter that, contrary to what surname regionalisation maps have suggested in the literature, not all resulting clusters are equally probable to occur within the data.

6.2.2.1 Consensus Clustering

Consensus clustering, first proposed by Monti *et al.* (2003), is a relatively new method for class discovery. It is premised on the idea that items consistently grouped together are more likely to be more similar than those appearing in the same group less frequently (Simpson *et al.* 2010). The method is designed to increase the stability of the final cluster outcomes by taking the consensus of multiple runs of a single cluster algorithm. Simpson *et al.* (2010) have provided an extension to this approach, called merged consensus clustering, by enabling the cluster assignments to be the product of multiple runs of multiple algorithms. By merging the results from different algorithms it is thought that the confidence in the result will increase because the limitations of one clustering algorithm will be offset by the strengths of another. For example Ward's hierarchical clustering is sensitive to outliers in the data, but offers a stable solution over-all in terms of consistency of cluster outcome; by contrast, the over-all arrangement of *K*-means clusters is relatively unstable, but the solutions are less sensitive to outliers. In addition to the increased stability of the results, consensus clustering can provide a range of metrics to help inform the optimum number of clusters as well as the robustness of the resulting cluster outcome- in terms of its structure and the membership of individual clusters.

Before undertaking the merged consensus clustering procedure, the user has to select the clustering algorithms to be used. Theoretically there is no limit on the number of algorithms that contribute to the final result aside from the practical constraints related to computation time and the degree to which the result will actually improve. Some of the most popular data classification methods in this context are Ward's hierarchical clustering, *K*-means, partitioning around medoids (PAM), self-organising maps (SOMs) and model-based Bayesian clustering. The algorithms selected for this study reflect those used in Chapter 5 and are listed in Section 6.2.3 below. Table 6-1

shows the definitions of the variables used – the latter are adapted from Monti *et al.* (2003) to make them more applicable in this context.

Table 6-1: Variables and definitions used in merged consensus clustering. Adapted from Monti *et al.* (2003).

Symbol	Description
$\boldsymbol{D = \{e_1, \dots, e_N\}}$	Data, in this case surname distance matrix with spatial units (e_i's) to be clustered.
$\boldsymbol{N}$	The number of spatial units (or number of rows) in distance matrix.
$\boldsymbol{P = \{P_1, \dots, P_K\}}$	Partition of D into K clusters.
$\boldsymbol{K, K_{max}}$	Number of clusters, maximum number of clusters.
$\boldsymbol{N_k}$	Number of items in cluster k.
$\boldsymbol{H}$	Number of resampling iterations.
$\boldsymbol{D^{(h)}}$	Dataset obtained by resampling D (h-th iteration).
$\boldsymbol{M, M^{(h)}}$	Connectivity matrix, corresponding to h-th iteration.
$\boldsymbol{\mathcal{M}, \mathcal{M}^{(K)}}$	Consensus matrix, corresponding to K clusters.
$\boldsymbol{I^{(h)}}$	$N \times N$ indicator matrix.

Consensus clustering first samples the complete dataset D to create a new subset $D^{(h)}$ before clustering using the specified algorithm(s). The sampling (using methods such as bootstrapping) and clustering are repeated multiple times in order to gauge sensitivity to repeat sampling from the total number N of randomly selected geographic units e_i. The results from each iteration are stored in a consensus matrix $\mathcal{M}$, which records for each possible pair of e_i the proportion of the clustering runs in which they are both clustered together. The consensus matrix is derived by taking the average over the connectivity matrices of every perturbed dataset (Monti *et al.* 2003). The entries to the matrix are defined in the following way:

$$M^{(h)}(i,j) = \begin{cases} 1 \text{ if } i \text{ and } j \text{ belong to the same cluster} \\ 0 \text{ otherwise} \end{cases} \tag{6.1}$$

$D^{(h)}$ is the ($N \times N$) connectivity matrix, required to keep track of the number of iterations in which both geographic units are selected by resampling, such that its (*i,j*)th entry is equal to 1 if both *i* and *j* are present in $D^{(h)}$, and 0 otherwise.

According to Monti *et al.* (2003) the consensus matrix $\mathcal{M}$ is the normalised sum of the connectivity matrices of all the perturbed datasets $\{D^{(h)}: h = 1,2, \dots, H\}$:

$$\mathcal{M}(i,j) = \frac{\sum_h M^{(h)}(i,j)}{\sum_h I^{(h)}(i,j)} \tag{6.2}$$

The *i,j*th entry in the consensus matrix records the number of times the two items have been assigned to the same cluster divided by the number of times both items have been selected (sampled). It therefore follows that a perfect consensus result would produce a matrix containing only 0s and 1s. $\mathcal{M}$ in essence provides a similarity measure to be used in further clustering or agglomerative hierarchical tree construction (Simpson *et al.* 2010).

To create a merged result a *merge matrix* provides a way of combining the outcomes of multiple methods by weighted averaging their respective consensus matrices (Simpson *et al.* 2010). The weighting can be adjusted to increase or decrease the influence of certain clustering methods. The advantage of this approach is that it mitigates the issues associated with the different classification properties in each of the algorithms discussed above.

Two types of clustering robustness measures can be calculated. The first relates to the cluster structure (called *cluster consensus m(k)*) and the second to the cluster membership (called *item consensus* $m_k(i)$). In regionalisation problems, the latter is especially useful because it enables the comparative visualisation of the geographic unit's cluster allocations alongside their summary measures of cluster robustness. As is often the case, a geographic unit may only be assigned to a particular cluster on the basis that all units have to be assigned to one of the set of clusters. Where allocations are marginal, there will be low confidence in the allocation and it can therefore be interpreted accordingly. Monti *et al.* (2003) first define I_k as the set of indices of items (geographic units in this case) belonging to cluster *k*. This can then be used to define the cluster's consensus as the average consensus index between all pairs of items belonging to the same cluster:

$$m(k) = \frac{1}{N_k(N_k - 1)/2} \sum_{\substack{i,j \in I_k \\ i<j}} \mathcal{M}(i,j)$$

(6.3).

The corresponding item consensus for each item e_i and each cluster k is defined as:

$$m_i(k) = \frac{1}{N_k - 1\{e_i \in I_k\}} \sum_{\substack{j \in i_k \\ j \neq i}} \mathcal{M}(i,j)$$

(6.4)

where 1 {cond} is the indicator function that is equal to 1 when cond is true and 0 when false. Item consensus $m_i(k)$ measures the average consensus index between e_i and all other items (geographic units) in cluster k. In the case of perfect consensus across all runs, the cluster consensus would be 1 for each cluster. As is demonstrated in the results section, this measure provides the level of confidence in the final result, expressed as a function of the number of times a geographic unit has been assigned to a particular cluster.

The use of multiple classification methods across a range of cluster values enables consensus clustering to provide a number of metrics to help inform the selection of the optimal number of clusters. Monti *et al.* (2003) state that the true number of clusters (k) can be estimated by finding the value of k at which there is the greatest change in the empirical cumulative density function (CDF) calculated from the consensus matrix $\mathcal{M}$ across a range of possible values of k. If the unique elements of $\mathcal{M}$ are placed in descending order, it is possible to define the CDF(c) over a range c=[0,1] using the following equation:

$$CDF(c) = \frac{\sum_{i<j} 1\{\mathcal{M}(i,j) \leq c\}}{N(N-1)/2}$$

(6.5).

It is then possible to calculate the area under the curve (AUC) of CDF as follows:

$$AUC = \sum_{i=2}^{m} [x_i - x_{i-1}] CDF(x_i)$$

(6.6)

where x_i is the current element of the CDF and m is the number of elements. If every iteration from the consensus clustering identifies the same groups then the $\mathcal{M}$

elements will be either 0 or 1, and thus $AUC = 1$. This provides the benchmark against which to compare the clustering results. One can experiment with the number of clusters into which to group the data, ranging from values between $K=2$ to K_{max} and compare their results with the benchmark $AUC=1$ result. The result with the number of groups that comes closest to this can therefore be considered the optimum number of clusters. To establish the best outcome the quantity ΔK is calculated, which is the change in AUC as K varies. The optimal k value is broadly considered to coincide with the peak in ΔK. Using Simpson *et al.*'s (2010) merged method the resulting consensus matrices (one from each cluster method used) from the optimal k are combined through weighted averaging. The merged matrix maintains the same properties as a consensus matrix and can therefore be used as a dissimilarity matrix for re-clustering.

6.2.3 IMPLEMENTATION

The Lasker Distance calculation remained unchanged from Equation 5.2 and produced a distance matrix from the 763 spatial units. This provided the input to the consensus clustering and MDS, which was implemented using R (R Development Core Team 2011); in particular the consensus clustering required the *clusterCons* package, developed by Simpson *et al.* (2010). The package is a new release and designed primarily for gene expression microarray data and this chapter is its first documented use in the context of population data.

A matrix of the Lasker Distances between all pairs of NUTS geographic units provided the input for the *clusterCons* package. Consensus clustering was performed using three different algorithms: *K*-means, partitioning around medoids (PAM) and Ward's hierarchical clustering. With the exception of PAM these were selected in the light of their success in the previous chapter. PAM is an additional clustering method to those applied in Chapter 5, outlined in Kaufman and Rousseeuw (1990), that differs from *K*-means by selecting at random a series of actual data points (rather than any point within Euclidean space) that are assigned to particular clusters based on their "nearness". Nearness is calculated using a pre-computed dissimilarity matrix

across all variables and data points within the data. PAM is less sensitive to outliers because positioning the centroid uses a median rather than the mean in the optimisation procedures. PAM minimises

$$V = \sum_{j=1}^{k} \sum_{i=1}^{k} \left| x_j - \mu_i \right| \tag{6.7}$$

where k is the number of clusters, and μ_i is the mean centroid of all the points x_i in cluster i. PAM has been selected here because its initial seed assignment, and therefore its final cluster outcome, is considered better than *K*-means. It was not utilised in the previous chapter because it is more computationally intensive rendering it inappropriate for the creation of regions with fine scale spatial units such as CAS Wards.

In order to select the most appropriate number of clusters (*K*) in which to group the geographic units, each of these algorithms were run using *K* values ranging between 5 and 45. For each value of *K*, subsampling was used to provide 200 selections for each algorithm in the consensus clustering. The results of this process produced a merged

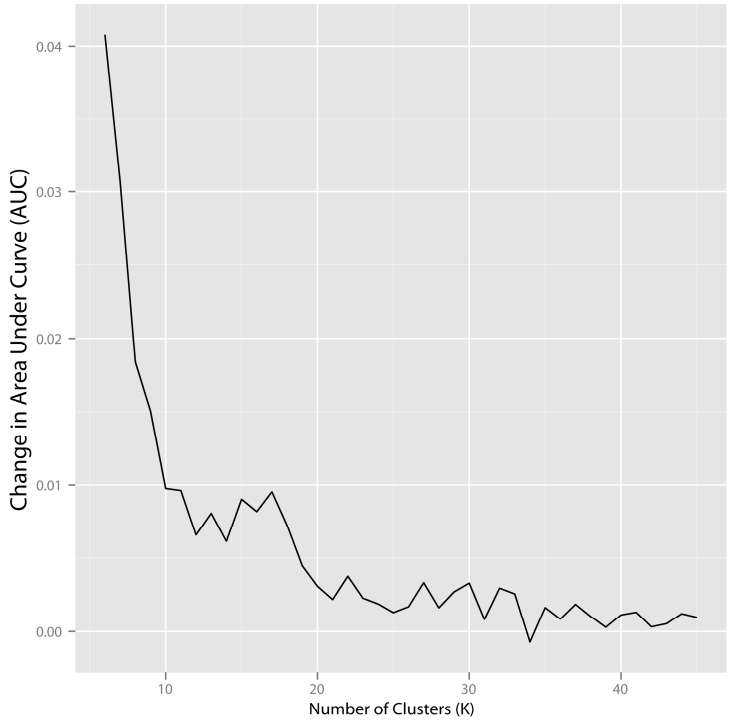


Figure 6-5: The delta *K* plot used to inform the decision to cluster the Lasker Distance matrix into 14 groups. It shows the change in AUC values are as calculated in Equation 6.4.

consensus matrix – an average of the three consensus matrices (one for each clustering methodology) – for each value of *K* (resulting in the creation of 40 matrices). The merged consensus matrices provided the basis for the ΔK calculations, the results of which are shown in Figure 6-5.

Figure 6-5 shows a dramatic decrease in ΔK values between *K*=5 and *K*=12, fluctuating between 12 and 20 before stabilising after *K*=21. On the basis of Monti *et al.*'s (2003) number of clusters criterion (outlined in Section 6.2.2) it was decided that 14 clusters presented the optimal outcome for the European data. This does not exceed a practical number of clusters for visualising regions in a choropleth map and it makes intuitive sense as it approximates the number of countries used in this analysis, and hence it is likely to capture the most significant interactions between countries. A number of results with more clusters were tested but, as predicted by Monti *et al.* (2003), random clusters were created with the consensus clustering methodology if the stopping criterion moves beyond the highest ΔKvalues. The results with *K*> 14 thus contained some questionable regional groupings within countries.

Figure 6-6 shows a box plot with the robustness values associated with the final cluster structures at 14 clusters (as outlined in Equation 6.4). In addition to the results from clustering the final merge matrix, those from the non-merged consensus clustering are also included for comparison. Consistent with preliminary research using different data (Cheshire *et al.* 2011), the merge matrix result produced higher median robustness values across all algorithms when compared with the non-merged results. Overall, based on Figure 6-6, it was thought that PAM on the merge matrix produced the most robust cluster structure. Although the PAM inter-quartile range was greater than that for Ward's algorithm, six of the 'Ward clusters' (nearly half) were classified as outliers. The membership robustness values were also highest, on average, for the PAM clustering result: these have been mapped alongside the final cluster outcome in Figure 6-7.

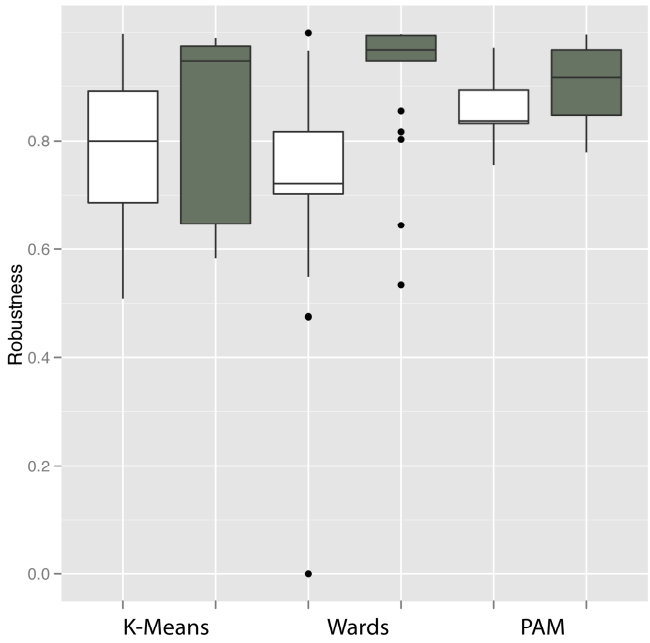


Figure 6-6: Box-plots showing the robustness values associated with the structures of each of the cluster outcomes. White boxes are produced from direct clustering of the distance matrix and grey boxes are produced from clustering the merged consensus matrix. It clearly shows that PAM provides the best solution in this instance.

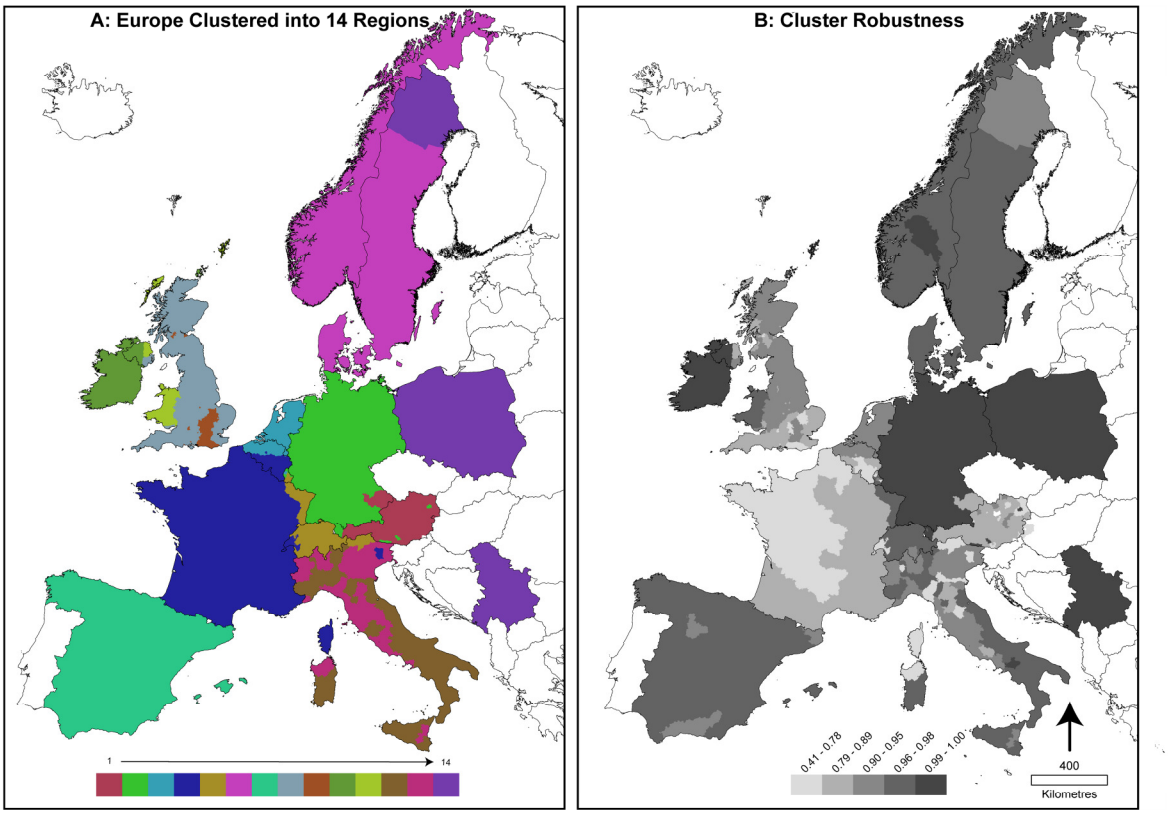


Figure 6-7: Maps showing the spatial distributions of each of the 14 cluster allocations (left) and their respective robustness values (right). On the left hand plot each cluster has been assigned a unique pattern.

6.2.4 MULTIDIMENSIONAL SCALING

In this chapter, following its success in the context of Great Britain, MDS in two and three dimensions is also used. MDS undertaken for greater than three dimensions had little impact on the positioning of the NUTS regions in relative space and becomes increasingly hard to visualise effectively in print. Results from the MDS are shown in two ways. Figure 6-8 shows a conventional plot of the results from two-dimensional MDS for each country, where each dot represents a NUTS region and each axis each of the two MDS dimensions. As with the MDS results in Section 5.4 the raw MDS values have been rescaled and to inform the colour values of the map in Figure 6-9. Each separate component is mapped onto one of these colours (Dim. 1= red, Dim. 2=green, Dim. 3= blue) before all three are combined into a single map in Figure 6-9D.

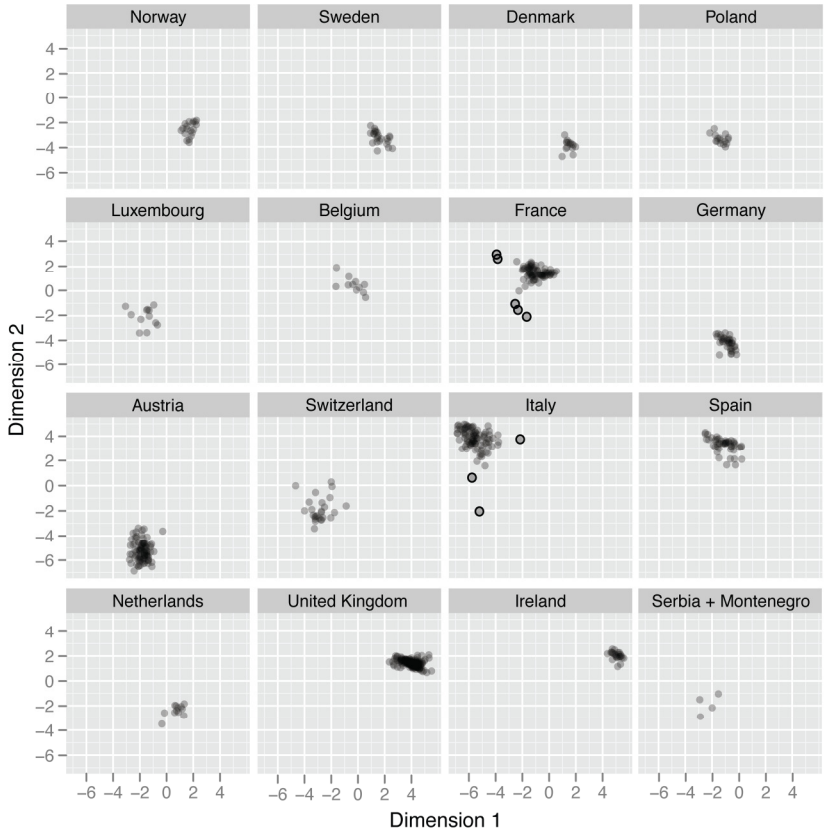


Figure 6-8: Plots illustrating the results of the 2-dimensional MDS analysis on the Lasker Distance matrix. Each country has been separated for ease of comparison and each point represents a NUTS region.

Finally, in order to measure the effect of 'isolation by distance' (defined in Section 2.3), Figure 6-10 (above) plots for each of 290,703 possible pairs of spatial units their geographic distance (measured as straight line distance in kilometres from the NUTs centroids) against their Lasker Distance in surname space (Equation 5.2). The same type of plot is also separately repeated for each country and shown in Figure 6-11.

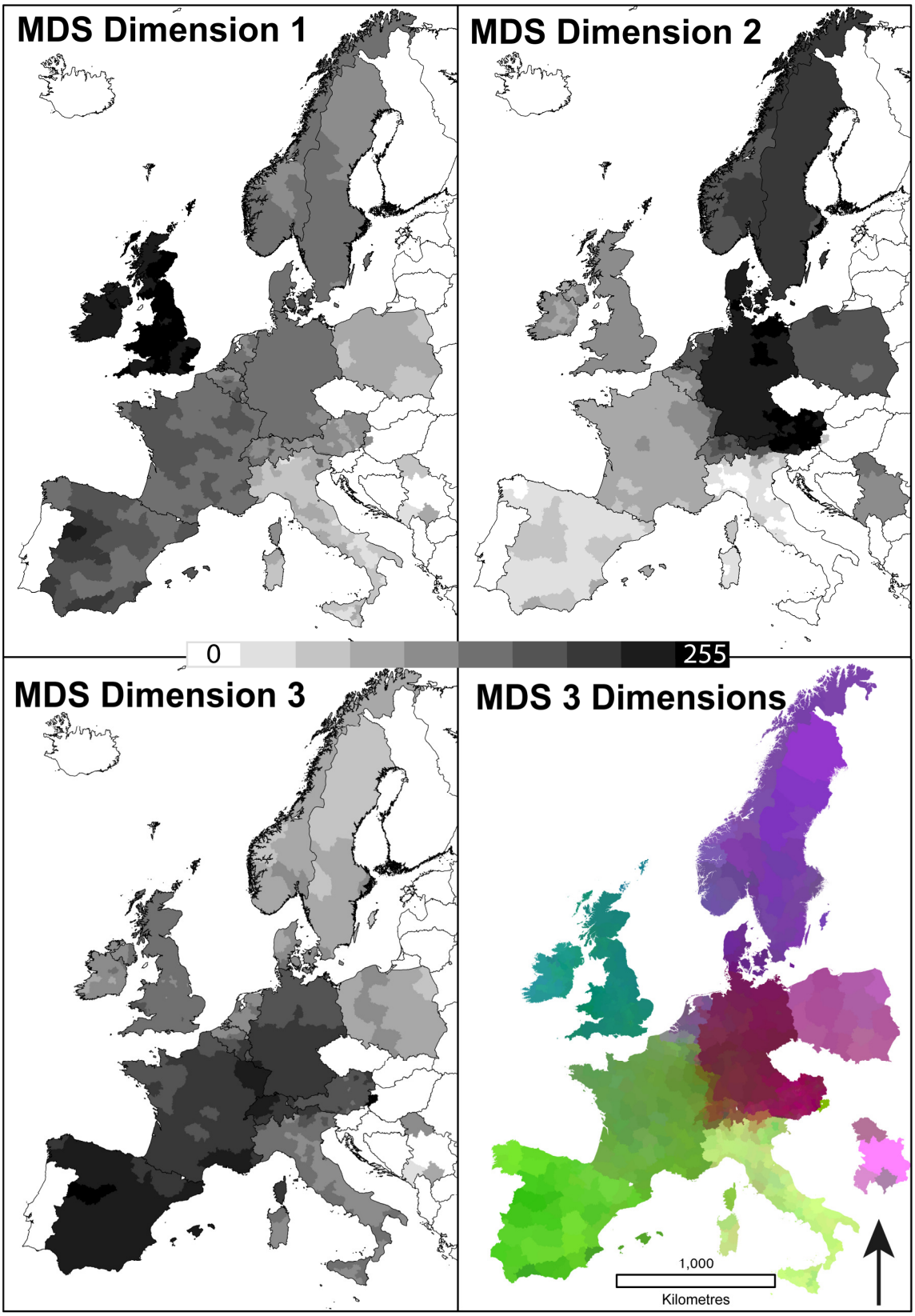


Figure 6-9: Maps showing the spatial distributions of each dimension produced from the 3 dimensional MDS. Each dimension has been rescaled to a value of between 0 and 255 to facilitate the creation of RGB colours.

6.2.5 RESULTS

This section presents the key results of the analysis presented above with the general objective of describing the geographical patterns found and offering some insights into the performance of the classification and visualisation methods used. The specific methodological aspects derived from these results will be discussed in the next section.

6.2.5.1 Isolation by distance

The scatterplot in Figure 6-10 hints at a relationship between Lasker Distance and

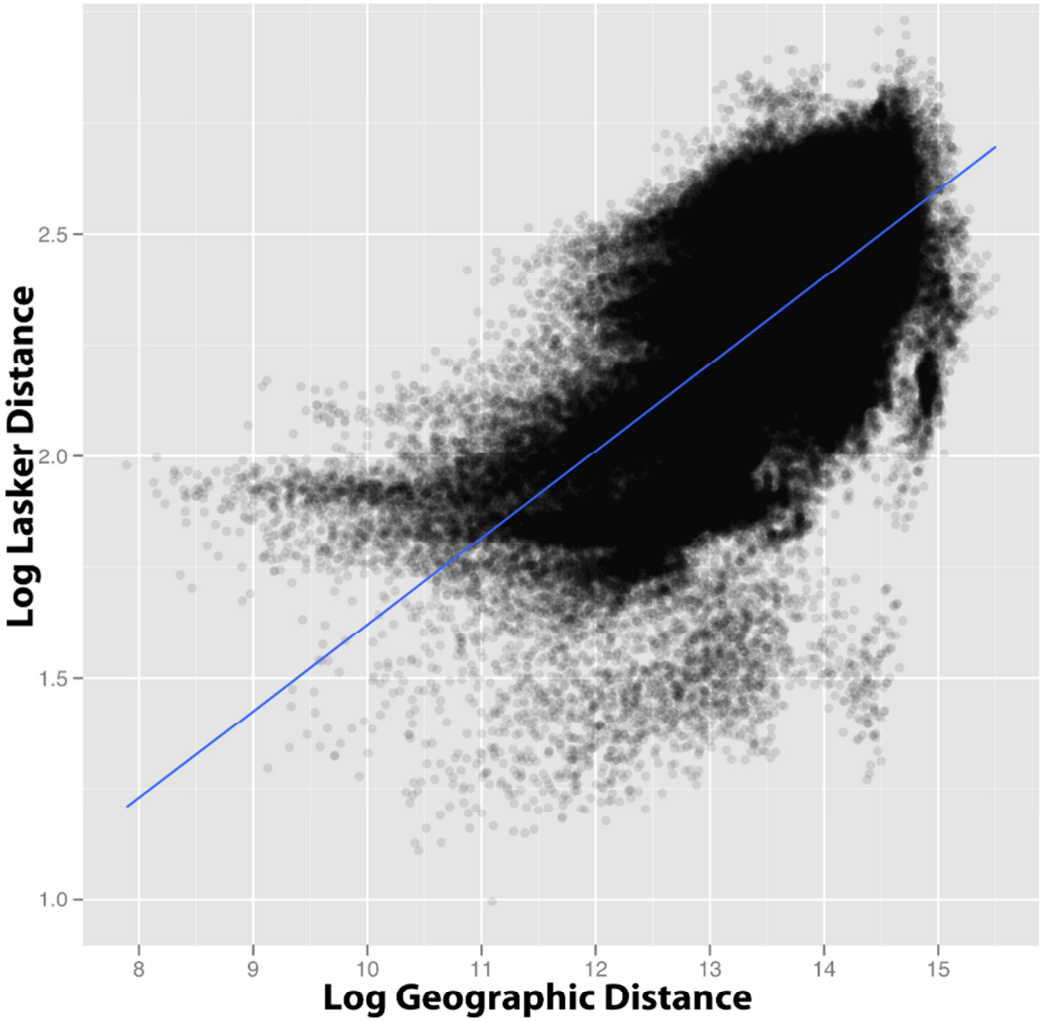


Figure 6-10: A plot showing the relationships between the Lasker Distance and geographic distance. Taking the log of each axis creates a greater spread of points in the plot window. Every possible region-pair is represented.

geographic distance across Europe, although the strength of this relationship may be less strong than could have been expected from general knowledge. This can largely be attributed to the fact that straight-line distance fails to reflect well-known physical barriers to movement, such as coastlines and mountain ranges that facilitate or impede movement. The mean Lasker Distance across Europe is 10.45 with the maximum value (19.68) occurring between Northern Ireland and southern Italy, hinting at a measure of isonymy with a low dispersion across Europe compared to geographic distances.

At the country level, the relationship between surname and geographical distance presents some interesting and particular national trends, as shown in Figure 6-11. Multilingual countries, such as Belgium and Switzerland, unsurprisingly show the strongest relationship between geographic distance and differences in the surname composition of its regions. Counter-intuitively perhaps, the plot for Norway suggests that surname diversity increases with proximity. This is most likely due to the greater surname diversity (resulting from domestic and international migration) in urban areas that are close to one other in the southwest of the country. This diversity appears to be sufficiently strong and in close proximity, managing to offset the more distant but more homogenous rural areas. In countries such as Denmark straight-line distance does not reflect actual population interaction because it has so many islands. Moreover, the plots in Figure 6-11 provide an important indication of the sub-national interactions between distance and surname diversity.

6.2.5.2 Consensus Clustering

The clustering results shown in Figure 6-8A conform to many well-known national and linguistic divisions across Europe, and most notably, follow linguistic or historical political boundaries, in some cases reflecting the effects of contemporary global migration to large urban areas.

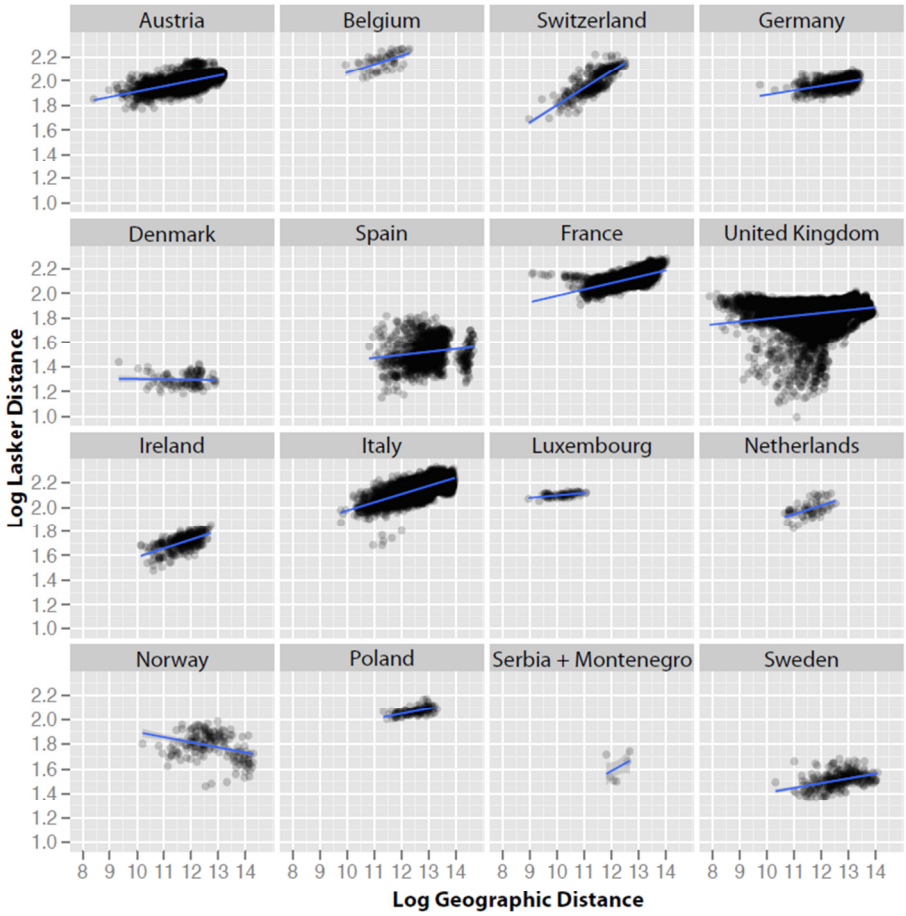


Figure 6-11: A plot showing the relationships between the Lasker Distance measures and geographic distance within each European country studied here. Every possible region-pair is represented.

The clusters generally follow national borders, with some interesting exceptions relating to multilingual countries and those with unique regional patterns. Large parts of Switzerland have been allocated to the same cluster as the Alsace region in France, Southern Luxembourg and the Bolzano region in Northern Italy, denoting similar surname characteristics shared by these multilingual areas with links to German language heritage. The analysis has also split Belgium along linguistic lines, assigning Flanders to the same cluster as the Netherlands and Wallonia to the French cluster, with Brussels appearing as a French enclave within Wallonia.

Denmark, Norway and Sweden have been assigned to the same cluster except for one sparsely populated area of northern Sweden known to have commonalities with its Finnish neighbour. This particular area has been grouped together with more

"peripheral" countries such as Poland and Serbia, Montenegro and Kosovo. The robustness values associated with this area in Sweden are low, suggesting that it shares little in common with the countries included in this cluster, which is essentially a Polish cluster, with the former Yugoslavia region being associated with it due to its small size in relative terms (in effect an outlier in the same way as the aforementioned northern Sweden region).

Beyond contemporary political borders there are some interesting within country regionalisations that derive from the analysis. In the UK, historical linguistic regions such as Wales, and the Scottish Islands are clearly distinguishable. It is also interesting to see the urban corridor around London suggesting that the surname composition of these areas is much more diverse and hence disconnected from the national picture. This demonstrates the uniqueness in the surname composition of contemporary global migrants to the London area (see previous chapter). In the rest of the British Isles, Ireland (Eire) is grouped under a single cluster, which includes most of Northern Ireland, except for the eastern coast, perhaps reflecting the close migration and trade flows with Great Britain.

In France, the mainland except for the Alsace-Lorraine has been allocated to a single cluster that includes the island of Corsica and the Geneva region in Switzerland, as well as the Wallonia region in Belgium. Italy has been split in two clusters, with a northern and western cluster separated from the rest of the country. Spain solidly belongs to a single cluster, despite its strong multilingual characteristics (Mateos and Tucker 2008), perhaps because of its overall low surname diversity (Scapoli *et al.* 2007). Most of (reunified) Germany is allocated to a single cluster, while most of Austria belongs to a separate cluster, with some spillover regions between the two.

6.2.5.3 Multidimensional Scaling (MDS)

The results from the multidimensional scaling largely support the consensus clustering outcome. The 2-D MDS plots for individual countries, shown in Figure 6-9, provide an indication of the location of each of the spatial units in their

multidimensional surname space. Those countries that have largely homogenous surname distributions form very tight clusters, such as Germany, Ireland or Denmark. Others such as Switzerland, Luxembourg, France or Spain, show a greater degree of scatter, reflecting present or historic multilinguism. Of most interest are the outlier points for each of the countries. For example, the three highlighted points in Italy's distribution are spatial units on the island of Sardinia, and those highlighted in France represent the border region of Alsace-Lorraine.

Figure 6-9 provides the geographic context to the results of the MDS analysis and is, in many ways, much more informative as a result. The maps create a similar impression to those in Figure 5-14 in addition to some more subtle distinctions. For example MDS Dimension 3 suggests a rather strong north-south split within Germany that is not noticeable in the consensus clustering results or the three-colour map. Multi-lingual countries are also clearly identified in Figure 6-9, as well as some of the diversity within the Netherlands identified by Barrai *et al.* (2002). It is clear from Figure 6-9 that the European map has a number of abrupt transitions in its surname compositions. There are clear splits between the British Isles and the Continent, between Romance and Germanic languages, between Scandinavia and the rest of Europe, and between Poland and Germany. The latter abrupt transition is especially striking since the current Polish-German border only dates to 1945. Many of the distinctions are perhaps unsurprising but these maps show, for the first time, how abrupt boundaries across Europe can simply be captured by surname frequencies largely derived from telephone directories.

6.2.6 DISCUSSION

The fact that the outcomes from the two separate regionalisation techniques used in this chapter, consensus clustering and MDS, are in broad agreement with previous research (see Chapter 2) is encouraging and serves to endorse their use in geographic analysis of population structure. Clustering the merged consensus matrix provided a more consistent outcome than non-merged consensus clustering, which in turn was more reliable than clustering areas using a single algorithm. The method does not

obviate the need for the selection of a single algorithm to produce the final result, but it does provide some useful metrics upon which to base this decision. As Figure 6-7 demonstrates, the ability to map the cluster membership robustness of each spatial unit to its respective final cluster provides a powerful way of assessing the appropriateness of the outcome for each specific area.

A key flaw with conventional clustering routines is the requirement to assign every item to one of a limited set of clusters, since this may result in questionable cluster allocations. Using robustness measures, such 'weak' allocations can be identified and interpreted with an appropriate degree of caution. In addition the ΔK measure is useful for indicating the optimal number of clusters; but it should be noted that "optimal" in the quantitative sense, might not be optimal in the practical sense. If the outcomes were to be mapped, for example, there would be a limit on the number of cluster outcomes that can be readily discriminated by the map user. A substantial advantage of the methods presented here, and in Chapter 5, is in the visual outputs that they provide so this limitation should not be underestimated.

A final consideration relates to the opposite scenario where the ΔK measure indicates that a very low cluster number is optimal but the researchers may wish to identify a greater number of clusters to highlight diversity. In this case the desired clustering result can be shown alongside that which is optimal. Merged consensus clustering cannot therefore entirely remove the need for subjective guidance of cluster analysis, but it does provide measures up on which researchers can base their decisions.

The maps shown in Figure 6-9 further demonstrate the power of mapping MDS values in this context. The resulting impression of regionalisation is similar to that produced by the computationally more intensive consensus clustering with the additional advantage that less discrete phenomena such as isolation by distance is also shown.

The datasets used in this chapter contain information at the level of the individual for most countries, and therefore, they offer the potential for much finer-scale analysis

with bespoke spatial units than has been presented here for the 763 NUTS2/3 areas. These will be likely to create subtly different regions from the same input dataset. This effect is clearly seen if Figure 6-12 from Scapoli *et al.* (2007) is contrasted with Figure 6-7A above. For example, Scapoli *et al.* (2007) have clustered the entire region of Lorraine as part of the Franco-German border area using NUTS 2 regions, while the smaller geographical units presented in Figure 6-7 (NUTS 3) suggest that it is only those departments contiguous with the German border (and not with Belgium or in the interior) that fall into this category. In addition, as is outlined in more detail in Section 2.4, the use of larger spatial units (and therefore larger population aggregations) will increase the strength of any correlations. This, for example, forms part of the explanation as to why the relationship between Lasker Distance and geographic distance (shown in Figures 6-10 and 6-11) is weaker than has been suggested in previous research.

The issue of scale is partially resolved through the application and context of the surname research being undertaken. If, for example, surname analysis is used as a proxy for genetic information (see Sections 2.3 and 7.2) at the European level then fine scale analysis may be unnecessary since most traits are only noticeable at coarse granularity (Cavalli-Sforza 2000). That said, as the previous chapter demonstrates, the use of smaller units of analysis will still preserve the geographically extensive trends if these are legitimate and not just artefacts of the spatial units used. A major advantage of smaller spatial units is their ability to highlight detail, such as that arising out of more recent migration events. This may be especially useful in the context of understanding segregation in global cities such as London, Paris and other large European cities. Whilst such fine-scale analysis would not be practical at a European level, it could nevertheless be investigated within each of the 14 or so groupings created in this study in order to identify the dynamics within each of these surname sub-regions.

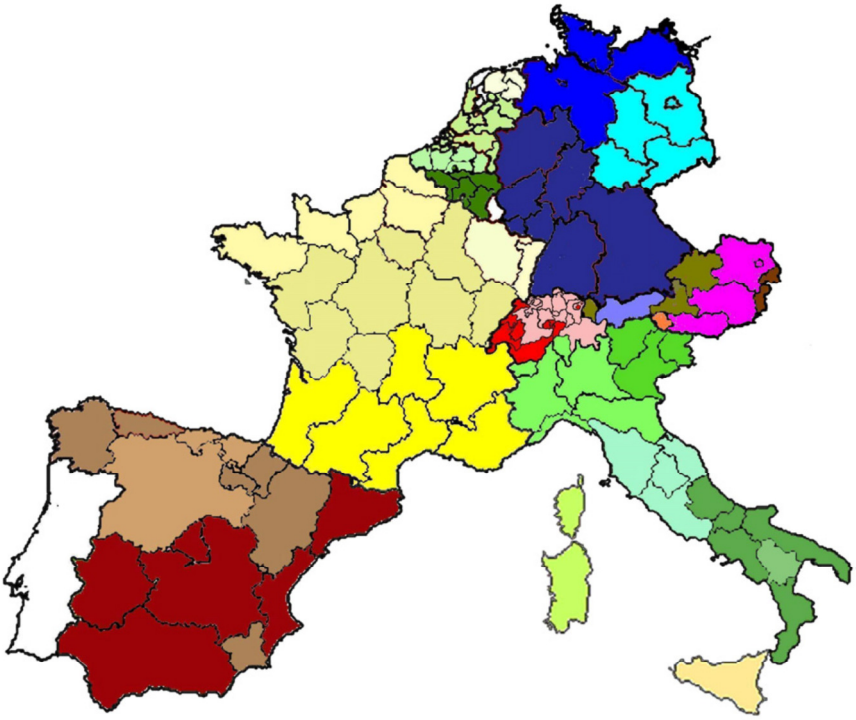


Figure 6-12: Scapoli *et al.*'s (2007) European Surname regions. Note the absence of the British Isles and Scandinavia and the relatively coarse spatial units used. Source: Scapoli *et al.* (2007: 47).

In summary, this chapter has demonstrated that there is a clear structure to the European population that can be discerned from the spatial distributions of surnames. Many of the regions identified are clearly linguistic, and perhaps unsurprising, but there are other some subtle cultural transitions identifiable as well. The Lasker Distance is a measure of relative (dis)similarity and therefore provides rates of change across Europe but with only the most significant transitions identified by clustering. The use of consensus clustering, its first implementation in this context, has provided additional information related to the robustness of the resulting grouping and also a useful range of metrics to help inform the optimal number of clusters. This information was necessary given the greater incompleteness of the data (in comparison to the previous chapter) and relative lack of trans-national knowledge surrounding *a priori* cultural, linguistic groupings within or between the 16 countries studied. Such considerations appeared less important for the MDS because,

as before, it has proved an effective visual measure of both the gradual and abrupt transitions in European surname geography.

6.3 GENERAL CONCLUSIONS ON THE REGIONALISATION OF SURNAMES

The two previous chapters have provided insights into the regional geography of surnames at a variety of scales (subnational, national and continental). The depth provided by Chapter 5 and breadth provided by Chapter 6 of this analysis is unprecedented in the context of surname research and this next section seeks to draw together some common themes pertaining to the regionalisation of surnames. The merits of the general classification approach taken in both chapters are first discussed before addressing the ways in which it can be improved and applied in further research.

Viewed in the context of the regionalisation tradition outlined in Section 5.1, the general approach here is refreshing because the data are allowed to speak for themselves, independent of any position or interest. None of the surnames have been filtered based on frequency or spelling and no weighting has been built into the classification to account for known historical similarities and differences. The similarities between spatial units have been calculated through a simple (dis)similarity metric in the form of the Lasker Distance. This has provided a consistent basis on which to undertake regionalisation at a range of scales and time periods. The isonymy measure, on which the Lasker Distance is based, remains one of the most widely used in names research and, as has been demonstrated in this thesis, can be easily applied to large datasets using straightforward computational procedures. This renders the isonymy measure a useful and indeed compelling technique for studies in regional geography.

Despite the applications of multiple classification methods on the Lasker Distance matrix, the results for each dataset have been surprisingly consistent, providing reassurance that genuine patterns are being represented. That this is the case, in spite of the almost inevitable noise inherent in some of the data, offers compelling evidence that there is no need for extensive editing of the data (through combining surnames with similar spelling, for example) based on subjective knowledge. Whilst intelligent sampling, such as the removal of migrant surnames, may enhance some of

the patterns this may risk providing only a partial picture of a ubiquitous cultural phenomenon. Such an approach also requires the definition of "migrant" surnames, for example, which in a country as diverse as Great Britain is simply not practical. In addition, it is also the case that variability caused by (inter)national migrant surnames is a characteristic of some areas, such as growing cities and central European countries, making it hard to justify migrant surnames in this context. The consistency of the clustering outcome also provides reassurance that combining a number of large and disparate datasets, each with their own levels of uncertainty into a single dataset for Europe is acceptable in this context. This is because, as demonstrated in Chapter 3, surnames have remarkably consistent relative proportions throughout population datasets, such as telephone directories, and can therefore be used to make scalable conclusions applicable to the total target population.

Despite the compelling regions evident in Figures 5-16 and 6-7 there are opportunities to improve these classifications. It is clear that some avenues will be more fruitful than others. For example, a common discussion in the regionalisation literature is whether geographical proximity or contiguity should form part of the final classification. In the context of surnames this is likely to be unsatisfactory because it prevents the creation of multiple geographically separated regions that nevertheless share common characteristics. For example, this applies to similar regions that have developed as a result of migratory processes between areas, such as the example of Cornish economic migrants moving to Middlesbrough in the 19th Century (see Longley *et al.* 2007), or the migration of Scots to Corby in England's Midlands described in Section 6.1.1. It also limits the types of diffusion processes responsible for the transmission of surnames to those requiring contiguous units. In reality it is likely to be the results of a number of processes (see Haggett 1994 for review). The regionalisation methods that have been set out here do not require boundary data: spatial attributes of the data are only utilised in the visualisation (mapping) and interpretation phases of the analysis and the identified regions have emerged without specifying contiguity constraints. Whilst contiguity is important in some applications (for example, when partitioning space for administrative purposes), in this context Johnston's (1970) enduring view that

"*regionalizing with contiguity constraints over simplifies and operates against efficient hypothesis testing. There is no basis in geographical theory...for the adjacency requirement*" (1970: 295) is most appropriate.

An investigation into the impacts of the size, concentration and distribution of the populations contained within the input spatial units would be a more useful area of research. Preliminary results for this research demonstrated the effect of the level of aggregation (size of spatial unit) and prompted the use of multiple levels of NUTS units with the European data. Each spatial unit is assigned an equal weighting in the analysis regardless of the size of its population. If a small population partitioned many times is represented by more spatial units than a large population partitioned a few times it will have a bigger impact on the final outcome through greater influence on the original dissimilarity matrix. The possible result is an impression of increased diversity in some areas over others. In the European case, a country's influence on the analysis is therefore based on the number of spatial units it has rather than the size of its population. This could lead to the potential for greater apparent diversity in countries tessellated into large numbers of spatial units but with fairly uniform surname compositions. The use of merged consensus clustering helped to mitigate some of these effects, in addition to minimising the impact of outliers in the cluster analysis. Future work as outlined in Section 7.3 should seek to establish a number of heuristics around which to base a suitable weighting methodology to account for the varying populations in each spatial unit across Europe.

Having established the empirical basis and validity of the regions identified here, this research might be used as a basis for further study of the clustering of surnames or areas in studies of regionalisation and population dynamics. Further temporal analysis of surname regions would identify whether the processes behind the creation of the observed discontinuities in surname structure are accelerating or attenuating, with consequential implications for the study of segregation within and between areas. This aim could be quite straightforward to achieve using the entirely computational approaches outlined above, but is, at present, limited by the availability of high-quality georeferenced population data.

In conclusion the creation and application of regional classifications is a rich field of study. The previous two chapters have demonstrated the utility of an inductive approach that takes advantage of recent technological advances to undertake intensive spatial analysis of surnames contained within complete population registers, in the case of Chapter 5, and telephone directories, in the case of Chapter 6. The concern has been the classification of places based on their surname compositions at a variety of scales. No previous research has assessed the appropriateness of the range of classification methods nor applied them to the depth and breadth that has been shown above. As a result, the great potential in the use of surnames for unearthing generalised population characteristics has been demonstrated. It is hoped that the surname regions produced can be used to inform future research in a range of disciplines.

7 METHODOLOGICAL CONTRIBUTIONS, APPLICATIONS AND FUTURE RESEARCH PROSPECTS

The results and analysis from the previous chapters provide ample evidence that surnames can make an important contribution towards understanding *people* and *places* – the former through analysis of the spatial distributions of people's names, and the latter based on areal surname compositions. The underlying concepts, methods and techniques that have been used to demonstrate this are of enduring importance in human geography, but the breadth and depth of their application to surname geography in this thesis is an important contribution towards understanding people and places, scale and aggregation, statics and dynamics. As Chapter 2 makes clear, previous attempts at the spatial analysis of surname data have been tentative or preliminary in terms of both the data utilised and the methods employed. This has inevitably placed limits on the nature and extent of generalisations that can be drawn, and limits the value of the inferences that can be made. It is in the context of these innovative, robust and analytical practices, allied to the exceptional content and coverage of the datasets used, that this next chapter seeks to consolidate the achievements to date, and identify a path for future developments. It will first outline the methodological contribution of this thesis to surnames research and spatial analysis more generally. It will then go on to demonstrate the relevance of the research undertaken here for population studies, including population genetics, before suggesting potential avenues for future research.

7.1 METHODOLOGICAL CONTRIBUTION

The methods applied and developed over the course of this thesis offer several insights into the spatial analysis of surnames and population data more widely. These approaches have been informed by a range of disciplines, principally quantitative geography and population genetics. Previous studies have not considered, in depth, the sensitivity of their outputs to choice of methodological approaches. The following section seeks to address this shortcoming by considering both the technical and conceptual insights offered by this thesis.

7.1.1 TECHNICAL

The previous chapters contain an in depth appraisal of a wide range of techniques able to capture and visualise the spatial distributions of surnames. In addition, a range of tools have been used to produce entirely automated, replicable classifications, using very large datasets. Even in comparison with geodemographics, a data intensive field in which computational data reduction techniques are the norm, the numbers used here are impressive: for example, the Lasker Distance matrix used to create the CAS Wards level regionalisation contained over 10 times as many data points (110,250,000) (see Section 5.4.4) than were required to create the UK Output Area Classification (OAC) (Vickers and Rees 2007). This capacity to handle large volumes of data has significantly increased the potential of quantitative methods in geography, and has led to the development of an entirely automated means of classifying the spatial distributions of surnames both individually and as regional aggregations. The outcomes are not definitive (and indeed are unlikely ever to be) but offer, for the first time, a standardised set of heuristics around which to draw inferences about people or areas from surname data. They also provide a consistent methodology on which to base future analyses as new datasets become available.

The use of kernel density estimation (KDE) to establish areas of spatial concentration is not new to spatial analysis, but this is its first application to surnames research. For the first time it has been possible to assign probable locales

of origin, areas of highest contemporary concentration, and areas containing specified proportions of population, to tens of thousands of surnames in Great Britain. This method is a marked improvement on previous methodological research, such as Manni *et al.*'s (2005) approach to identifying surname origins in the Netherlands using self-organising maps (SOMs). The most significant analytical advance is the entirely automated nature of the KDE method and its serial treatment of different surnames. This automation, for example, includes the recalculation of the bandwidth for each surname in order that it is sensitive to context, and the ability to change a variety of parameters to ensure the output is appropriate to the application. The SOM approach, in contrast, assigns origins to groups of surnames that require manual disaggregation at the end of the process; this would not be practical for the volume of surnames used here. In addition, the utility of both the continuous and discrete representations, provided by the KDE methodology, of the surname distributions offers the flexibility to alter the parameters, such as the population threshold, and to change the discrete representation of the data (in the form of a polygon derived from the contour line – see Figure 4-12 for example) without affecting the underlying continuous surface. An advantage of Manni *et al.*'s SOM approach, however, is its ability to identify surnames that share very similar points of origin and therefore could be different variants of the same root surname (Manni *et al.* 2005). Using KDE, this information would be obtainable but would require additional processing of the core areas encapsulated by the contour.

The outcomes from Chapter 4 go beyond the straightforward mapping of surname concentrations to provide a large number of metrics associated with their spatial distributions, such as areal extent, relative concentration, or relationship to place names. As Section 7.2.2 demonstrates, it is possible to mine such information to better select individuals or areas for sampling. Previously, KDE has been used in the context of characterising areas based on a point distribution, such as crime occurrences, or less frequently to assess the location of a fixed service provider, such as a school or doctor's surgery relative to its constituency of users. The data produced from the KDE approach can therefore be thought of as a generalised geodemographic classification, and the first implementation of KDE devised specifically for this purpose. The academic contribution of the methods and results

should be viewed not only as a technical advance in the context of previous surname research, but also as an applications advance within geodemographics more generally.

The opposite is the case for the use of rescaled multidimensional scaling (MDS) values to inform the map shown in Figures 5-14, 5-20 and 6-19. Similar approaches have been used in linguistics (see www.let.rug.nl/~kleiweg/L04/) and, by extension, the study of surnames, but not within geodemographics due to the subject area's preoccupation with assigning individuals to discrete categories. The maps provide a strong visual indication of the (dis)similarities between spatial units and show that some transitions in surname compositions can be gradual, while others are abrupt. The use of three variables to inform colour in RGB space is not new but its deployment with MDS here could stimulate useful extension in geodemographics. As Figure 7-1 shows, the script written for this thesis can been used as an alternative to visualising demographic differences in London. In this example a distance matrix has been produced from the same 41 variables that are used to create the OAC; and then, in the same way as the Lasker Distance matrix, had its dimensions reduced to three using MDS in order that it can be visualised in the RGB colour model. There are acknowledged limitations associated with perceptions of colour that may influence interpretations of the final result but these can be mitigated, albeit with a loss of information (if the colour models have fewer than three dimensions), with the use of perceptually uniform (meaning the magnitude of change in a colour reflects its visual importance) colour models, such as CieLab, shown in Figure 7-2. Despite these limitations MDS remains both an intuitive and effective method of demonstrating varying rates of change across space with a number of natural applications beyond surnames and linguistics.

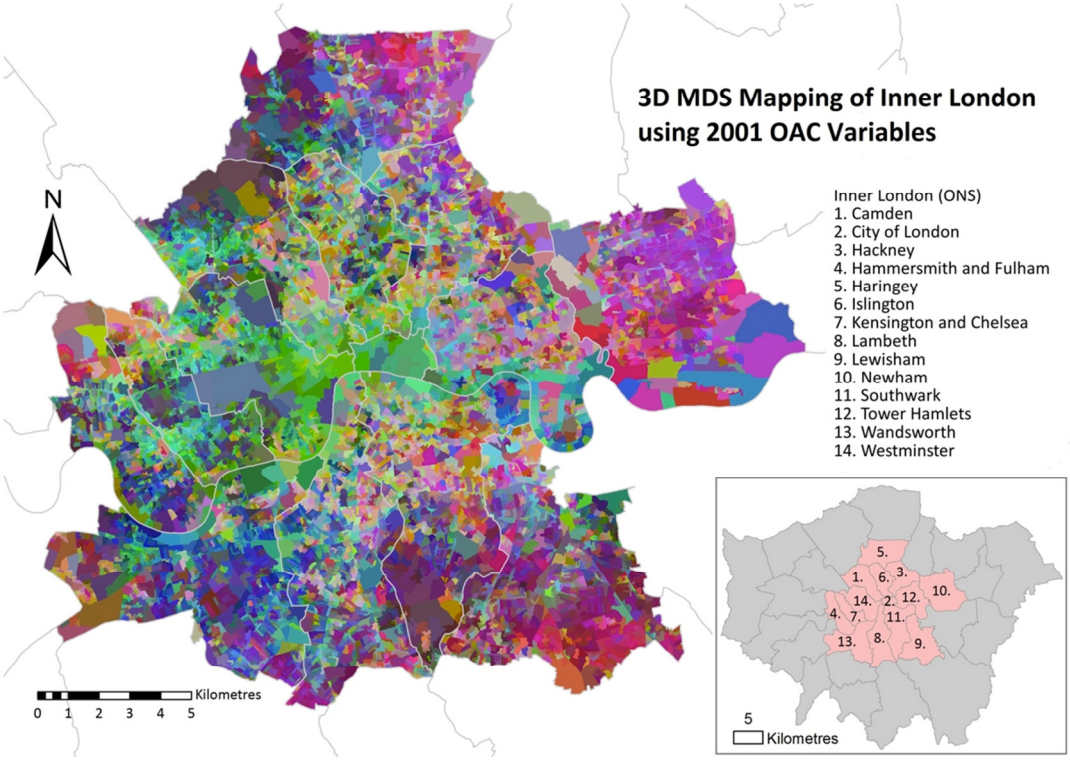


Figure 7-1: An example of using MDS to inform the colour values for a map of geodemographic characteristics. It was created from a distance matrix produced from the 41 variables used in the Output Area Classification (OAC). It provides an effective means of identifying areas of similarity and difference within selected inner London Boroughs without having to assign OAs to distinct clusters. Source: danieljlewis.org

One of the most promising technical contributions emerging from this thesis is the application of consensus clustering to create the European regions shown in Section 6.2. Its use was prompted by relatively little consideration for the impacts of uncertainty in conventional methods used by geodemographic classifications. There is increased uncertainty associated with the European data, in comparison to those for Great Britain, because of the number of data sources used and the limitations of a single level of aggregation. Techniques such as bootstrapping are frequently employed to gauge the robustness of cluster outcomes, for example in datasets with small sample sizes but large numbers of variables: see Monti *et al.* (2003). In reality, the computational intensity of consensus clustering would prevent its application at the CAS Ward level; however, as the limits to single chip processors are becoming apparent parallelised computing infrastructures, particularly graphics processing units (GPUs), are likely to reduce this limitation over the coming years (see Cheshire *et al.*

(2011), Adnan *et al.* (2010)). Consensus clustering provides important context to the regions produced such as the probability that they would be created in future runs of the algorithm. This is useful as it enables researchers to look at potentially dubious outcomes (based on their *a priori* knowledge) and establish if they are the product of forcing data into groups or if they depict genuine patterns in the data. Such information is becoming increasingly important as users are becoming used to analysing data interactively and on the fly; but this may be deceptive without sufficient feedback where results are sensitive to initial conditions or a predetermined number of groups (Adnan 2010). Enabling a degree of subjective interpretation is discussed in more detail in the following section.

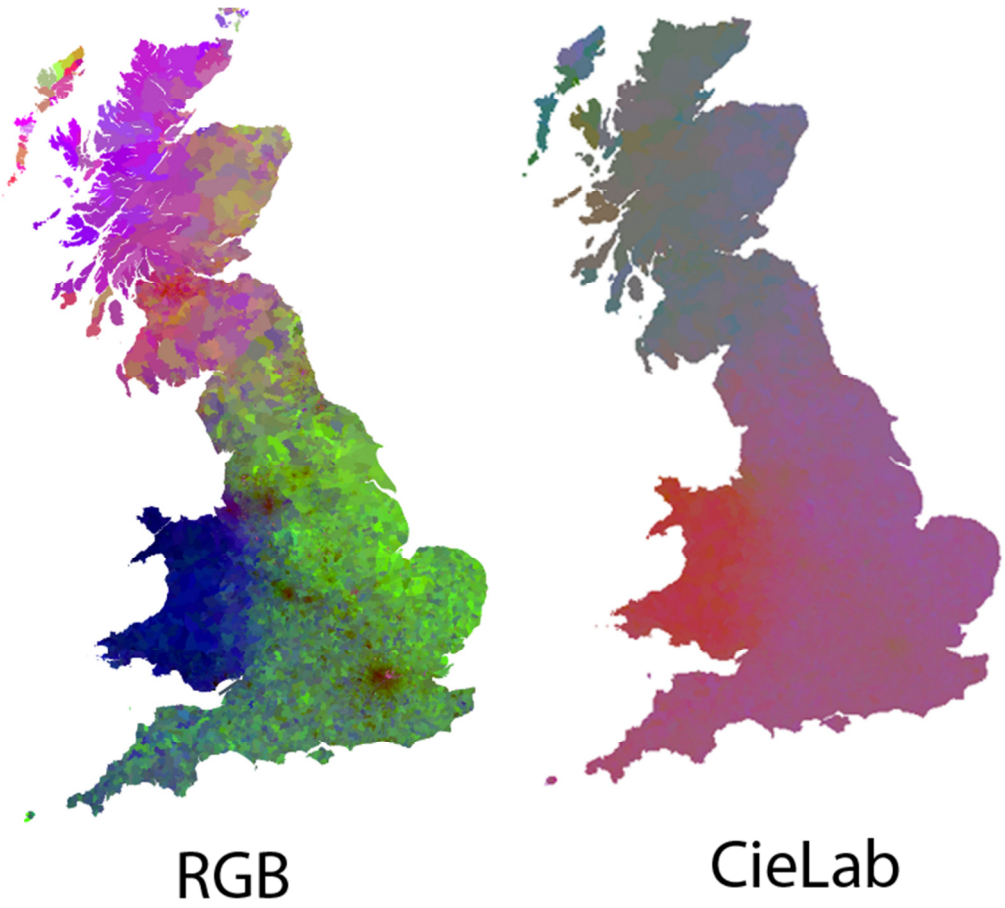


Figure 7-2: Comparisons between the red, green, blue (RGB) colour model and the CieLab model for mapping the MDS results from the Lasker Distance calculations. CieLab is designed to be perceptually uniform, but it conceals many of the more subtle distinctions in surname composition.

A final outcome from consensus clustering is that it could represent an improvement on previous approaches by offering increased consistency in the classifications from different clustering algorithms. This means that observed changes over time, for example, are more likely to be a product of the data rather than of the clustering algorithms used. This is useful in the context of surnames, but also in geodemographics more broadly as Figure 7-3 demonstrates. Consensus clustering has

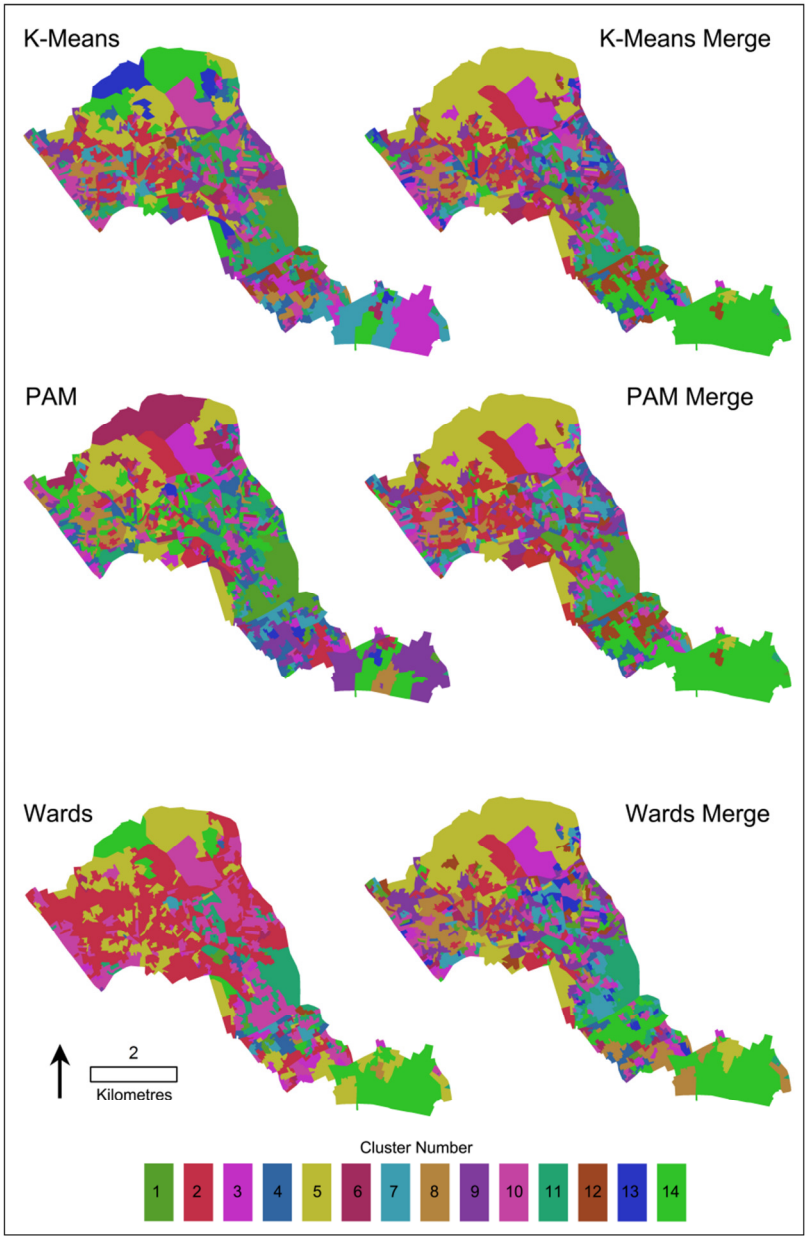


Figure 7-3: The mapped cluster outcomes from conventional clustering (left) and merged consensus clustering (right) of the 41 Output Area Classification (OAC) variables for two London boroughs (Southwark and the City of London). The latter shows a much more consistent outcome. This is useful, as temporal comparisons can be made in the knowledge that differences in the result are the product of changes in the data rather than the result of the cluster algorithms. See Cheshire *et al.* (2011) for more details.

therefore demonstrated its potential and is therefore worthy of further research to establish its strengths and, more importantly, limitations in this context.

7.1.2 CONCEPTUAL ISSUES

Aside from the technical insights provided by the methodologies employed in previous chapters, this research has encountered a number of conceptual issues worthy of further discussion. Some of what follows will be unsurprising in the context of previous surname research undertaken by population geneticists, whilst other aspects will be better known to quantitative geographers. The key concepts addressed in this section concern: the utility of inductive generalisation; the impact of scale and the level of aggregation in elemental units of this analysis; and the merits or limitations of continuous versus discrete interpretations of surname distributions. Implicit in much of this discussion is the extent to which process can be inferred from pattern and the way in which interpretations of pattern are affected by the concepts listed.

Much of this thesis is based on the premise that inductive approaches, when applied to comprehensive population datasets, are capable of creating valid and meaningful generalisations of surname distributions. This has indeed been demonstrated, with the results providing insights into the spatial behaviours of surnames that are in broad agreement with previous, more deductive, approaches. Such insights, at a range of scales, have been gleaned through automated methods that would have previously taken months of manual processing to obtain. Using the outcomes from this approach it is immediately possible to focus the geographic area over which researchers (or amateur family historians) might search for a particular surname, as a method of targeted sampling. Crucially, the automation of these approaches facilitates their straightforward application to new datasets as they become available. This is especially useful in the context of temporal research and will maximise the use of historical datasets without the constraints of manual interpretation.
The arguments in favour of an inductive approach to generalisation outlined in the previous paragraph are well known and are reinforced by the analysis in this thesis.

More interesting, however, are the influences of prior knowledge and desired application on the final outcome. Both the methodology to establish surname core areas and those to create surname regions intentionally require some user interaction. In Section 5.3 subjective inputs into an inductive classification were portrayed largely as negative features, but it has become increasingly clear through the methodological processes leading to the outcomes of this thesis that well-informed (by quantitative metrics and known distributions such as the Corby example (see Section 6.1.1)) interactions can improve the classification outcomes. This is especially evident when selecting the "optimal" number of regions in which to cluster the spatial units based on Lasker Distance. A good classification needs to be easily interpretable, thus placing an upper limit (20 in the case of the Great British classification) on the number of groups. In addition, it is always possible to cluster data into groups but a point will be reached when they will cease to be useful. On this basis the lineage between motivation and outcome needs to be established. The decision is also related to the scale of the analysis, as differences will be likely to decrease with smaller target areas; this is discussed in more detail below. Optimal in the statistical sense may not therefore coincide with optimal in terms of visualisation. The selection of the appropriate number of clusters within the data can never be entirely objective. It can, however, be informed by as many substantive, methodological and technical metrics as possible, such as those available from consensus clustering (see Chapter 6), and also in combination with techniques such as MDS. The importance of mapping the outcomes is also clear as it does much to establish whether, visually at least, the results conform to *a priori* expectations.

A further noticeable outcome from the inductive approaches used here is the way in which the final classifications can vary between the methods used. This relates partly to the different purpose of the methods with, for example, barrier algorithms designed only to detect the most abrupt transitions in the data, in contrast to hierarchical clustering, which is principally concerned with seeking homogeneity. Even when accounting for such differences and using a single dissimilarity measure, this work has shown the merits of combining interpretations from a number of methods to produce a more generalised and holistic impression derived from the classifications. Previous research in this context has only used single methods of

clustering and visualisation and therefore lacks the comprehensiveness that multiple methods provide.

The second concept identified above is the impact of scale on the impression of surname distributions. One of the most important influences of scale on both the creation of core areas of concentration and surname regions is the frame of reference it provides. For example, the differences seen between the regions of Europe in Figure 6-7 are greater than those of Great Britain in Figure 5-16. This is because the former is more extensive and has more variation within its global population. It is, however, also the case that at the finest scales, such as the urban areas shown in Figure 5-17, there can be greater differences than those measurable at the more generalised national level for Great Britain. This is demonstrated by the MDS plots in Figure 5-19 where the scatter in the plots produced by urban areas suggests larger differences within Government Office Regions (GORs) than between them, in contrast with the MDS plot in Figure 5-15; something that, in part, relates to the smaller spatial units used and their associated increase in internal homogeneity. More importantly, it also reflects the often abrupt local differences that can only be seen at the finest scales. In some cases these are the equivalent of crossing a national border in Europe due to the dominance of some groups in the surname distribution, for example Bangladeshis in East London adjacent to those with Anglo-Saxon surnames. This is illustrated in Figure 7-4 that shows the most popular surnames across a number of London Boroughs. It is the case that such differences are a largely urban phenomenon, in Britain at least, because the functions of towns and cities encourage and sustain the diversity required to produce abrupt transitions. This is a clear example of the relationship between pattern and process alluded to above. Rural areas are typically much less diverse and are captured by the gradual colour transitions shown in Figure 5-20. The outcome of the classification is therefore dependent on the variability of the underlying phenomenon of interest.

Based on the distinctions above, Figure 7-5 presents a crude representation of how the magnitude of expected variation is expected to change from the urban to the global scales. This effect could be exaggerated or suppressed depending on the size of the spatial unit so it needs to be viewed in the context of optimal variations in the levels of aggregation appropriate to the scale. As has been clearly demonstrated (Chapters 5 and 6), surname distributions produce a consistent picture across a range of spatial units, so having fine scale granularity at larger scales will not be as problematic as coarse granularity at smaller scales. The suggested changes in the magnitude of variation are illustrated by the mean Lasker Distance values. As Table 5-1 shows, the mean for Europe (10.45) is 60% larger than for the Great Britain calculation at Local Authority District level (6.31), which in turn is slightly smaller than the CAS Wards calculation for Great Britain.

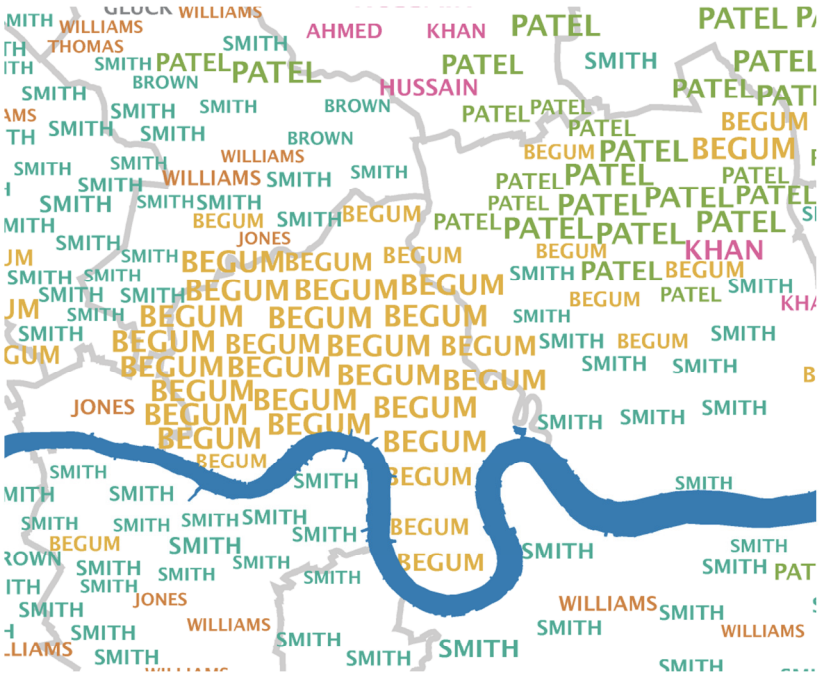


Figure 7-4: A map of Central London illustrating the most frequent surname in each Middle Super Output Area (MSOA). It demonstrates the dominance of certain surnames and the abrupt transitions from those of Bangladeshi origin (in orange), for example, and those from other cultures. For a full version of the map see Appendix 3 and names.mappinglondon.co.uk.

A further aspect of scale, and one that relates to Figure 7-5, is the point at which the magnitude of variation between spatial units is sufficiently great to be worth

partitioning into regions. As has been discussed previously, clustering algorithms will always create groups within the data regardless of how meaningful they are. In the case of creating regions this is a fundamental issue that can only be addressed by the prior knowledge of the researcher. Some clues will be provided as regions become increasingly fragmented and at the finest scales will probably only identify family groups. A further issue related to clustering is whether regions need to be nested (that is can be combined to create larger regions that still reflect surname composition) as the scale varies; this is something that would be straightforward to implement, if necessary, with sufficiently detailed data. Nesting would be useful if the regions are used as a unit of analysis (in common with census geographies, for example), as discussed in Section 7.2.3 below, but would impose unnecessary structure on the clustering outcome if they were created in isolation.

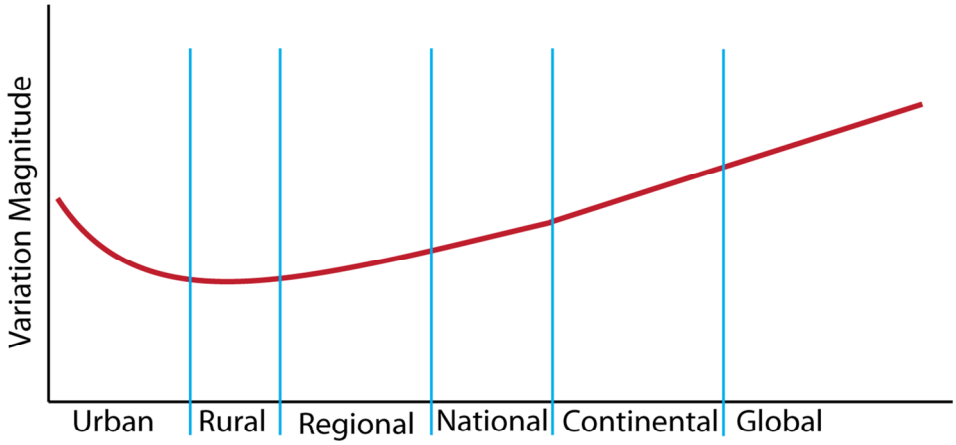


Figure 7-5: A hypothetical illustration of the impacts of different scales on the magnitude of variation between surname compositions. For this distribution to hold, the spatial units will need to be appropriate to the scale being viewed.

The final conceptual outcome concerns whether the analysis outlined in previous chapters has offered any insights into the best form of representation for surnames; that is are they best captured as discrete or continuous phenomena? Unsurprisingly, the answer to this question is dependent on context. In the case of identifying surname cores, the generalised impression given by a density surface is a useful visual aid and provides a direct comparison with many of the clinal conceptions. It also remains the case, as was discussed in Chapter 4, that surface based approaches minimise incompatibilities associated with inconsistent spatial units because all

analysis can be conducted on a regular grid. The ability to easily add a historical dimension makes the methodology presented here a powerful tool for comparing the spatial extent of population groups. In many cases, however, a more discrete interpretation (achieved using a contour line) reveals useful information, such as the spatial relationship between a place name and its associated toponymic surname.

In summary, it is clear that representations of the spatial distributions of surnames are affected by widely understood conceptual issues associated with spatial data. Where this thesis differs from previous research is by being the first to formalise them in the context of surname analysis. Despite many years of, albeit sparse, research into the spatial distributions of surnames, an awareness of scale, or an appraisal of the dominance of inductive approaches, has not been demonstrated. It is hoped that the discussion above provides firm conceptual foundations for future studies concerned with the spatial distributions of surnames.

7.2 APPLICATIONS OF SURNAMES TO POPULATION STRUCTURE

It is important to acknowledge that contemporary surname distributions are the outcome of generations of population movements that have taken place since surnames came into common parlance. They provide one of the few easily measurable and universal cultural markers, which, as this thesis has demonstrated, have unique spatial distributions that when combined create a clear regional geography measurable at a range of scales. On this basis alone, surnames offer important contextual information for a range of population characteristics. It is not the purpose of this section to comprehensively list the circumstances where surnames are relevant. Instead two examples are given; the first relates to indicators of population continuity and change and the second describes the potential importance of the results from this thesis to population genetics.

7.2.1 POPULATION CONTINUITY AND CHANGE

Population continuity is an appropriate term to refer to the amount of movement within a population, with those areas experiencing low migration considered as having greater continuity in comparison to those with high migration. Based on this definition, the "population" can refer to a group of people sharing a common surname, or a group sharing a common location. It is therefore possible to characterise a person as, for example, being part of a dynamic surname population (in the sense that their fellow bearers have migrated large distances) but a relatively static regional population or any other dynamic/ static combination. As the following discussion will demonstrate such characterisations of people and places are straightforward and insightful using the analysis undertaken for this thesis.

The methods outlined in Chapter 4 demonstrate a number of metrics to classify individual surname distributions. Based on such information it is possible to gauge the temporal dynamics of the distributions of long established surnames in Great Britain: some remain static whilst others have become more diffuse or shifted

concentrations towards urban areas. Such conclusions can only be drawn if the surname was classified as having any core area(s) at all, which is an important consideration if, for example, geographic ancestral information is to be obtained. With the exceptions of surnames that are too dispersed to have a defined core area (reasons for this are explored in Section 4.2) it is possible to establish the distance from a surname's core area of concentration that any individual has migrated. This information is useful if, for example, possible genealogical links are being traced between individuals (mostly males) sharing the same surname.

The surname cores calculated for 1881 can therefore be treated as the baseline surname geography to be compared with later datasets, in order that a detailed picture of gradual population movements over the past century can be constructed. It follows that inferences can be drawn about migration over the past century (within the limits of the data for Great Britain) and beyond (assuming that contemporary concentrations reflect surname origins). If more historical data becomes available, such as the 1841 census, it would, for example, be possible to gauge the impact of migration arising because of the industrial revolution on the underlying structure of the British population. Given the over-all similarities between the 1881 and 2001 comparisons, in spite of unprecedented population movements and influx of migrants, it would be surprising if the magnitude of change were greater between 1841 and 1881.

Preliminary work has already extended the ideas above by estimating the number of individuals who remain where their surname is most concentrated. One contention is that where there are large numbers of individuals living within their surname's area of highest concentration, there is likely to be greater continuity of community structure than those with lower percentages. Comparisons between the contemporary population and 19th Century provides a direct indication of the proportion of surname lineages that have remained centred upon their original core areas and the degree to which they have been subject to drift in the context of national trends and local susceptibility to these trends. The use of 1881 data also removes the impacts that recent migrations have had on the relative densities of surnames because "new" migrant surnames are not included. Initial findings suggest that the proportions of

surnames existing in their 1881 and 2001 core areas remains stable at around 2.7 million individuals (using the 95% contour threshold).

Establishing the core areas of concentration for a surname provides an indication of both continuity and change in population structure. Areas of recent change through influx of migrants (bearing names imported from abroad) will have high proportions of their populations living where their surnames are most concentrated in the UK. Such concentrations are a more recent phenomenon and are likely to be ephemeral due to the dispersal of bearers of those surnames, and their offspring, to other areas of Great Britain. More persistent concentrations exist in areas away from urban centres, such as in Cornwall, where a large proportion of the population, and their ancestors, appear not to have moved. As will be explored below, this information is useful to population geneticists.

In addition to the range of population characteristics, past and present, revealed by the spatial distribution of an individual's surname, it is also possible to derive information associated with the population characteristics of particular places. This can take two forms. The first is the amount of mixing within an area; and the second is the amount of mixing between areas. In the case of the first, a high diversity of surnames would suggest that a regional population has been subject to significant flows of migration, whilst lower diversities tend to indicate more stable populations of regions that are less central to migratory movements. The second is subtly different in that it relates to between area similarities and can be thought of as a measure of interaction. This is demonstrated by the plots of Lasker Distance against geographic distance in Figure 6-10. It is clear that Lasker Distance decreases with proximity, suggesting that the populations of nearer areas interact more to produce similar (but not necessarily less diverse) surname groups. If this relationship were absent the mapped cluster outcomes and MDS would produce less spatially contiguous groups.

Areas or regions that contain populations not subject to the normal ebbs and flows of movement will appear as anomalies in the relationships described in the previous paragraph. The largest exception to the effect of distance appears to be national and

linguistic borders, with many coincident with major transitions in surname compositions equivalent to hundreds of kilometres of change within a country. In addition, MDS provides indications of the porosity of jurisdictional boundaries. For example, the Scottish border appears more abrupt than the Welsh equivalent, and in the context of mainland Europe, central countries appear more porous than those on the periphery. This is largely a product of both geopolitics and language, but can also be used to provide additional context for studies of European mobility both in terms of historical migration and the likelihood of migration to some areas rather than others. Within areas of homogeneity, abrupt transitions can represent large-scale migration for specific, often economic, purposes. Examples of this, such as the Scottish migration to the town of Corby, differ in terms of their impact on the overall trends in surname compositions when compared with national borders. In such cases the surnames continue to change gradually through the process of isolation by distance around these areas, whilst national borders often represent a step change across their entire length. This information can be extrapolated to predict future levels of population interaction.

The distinction between localised migratory events affecting surname compositions and the broader impacts of both distance and linguistic (and national) borders serves to demonstrate the two levels of population structure that have emerged from the analysis of surnames. The underlying structure of gradual transitions over distance shows continuity in the population over many generations. It is the net result of the unique distributions of individual surnames characterised in Chapter 4. Changing populations tend to produce abrupt and localised transitions causing minor disruptions to the overwhelming impression of continuity shown in many of the outputs from this thesis. The crucial aspect of these areas is that they appear to be easily identifiable as distinct and contiguous clusters. If desirable, this facilitates their removal from the general trends to be treated as noise, or subject to separate study. This structure alluded to in Section 5.4 is of interest to many, especially population geneticists, who may wish to study areas with stable populations over time or, alternatively, those subject to a high amount of flux and mixing.

This work has not resorted to sampling to increase the impacts of particular surnames or groups; instead it has simply sought to represent near-complete population datasets. The results of thesis can therefore make a powerful statement about population mobility in the past century or so. If random migration were commonplace the patterns and distinctions highlighted in the results above would be non-existent or, at the very least, less obvious. Instead it would appear that surnames, and therefore their bearers, are surprisingly inert and presumably must have been over many generations. Those that have moved, or arrived from abroad, have done so predictably to urban areas, which has enabled the preservation of an underlying structure of general trends overlain with more abrupt transitions coincident with towns and cities.

7.2.2 SURNAMES AND SAMPLING FOR POPULATION GENETICS

Multiple references have been made throughout this thesis to the close relationship between surnames and population genetics. The link between them, outlined in Section 2.3, has provided sufficient motivation for the majority of surnames research to be undertaken from this perspective. Population geneticists, in this context, are interested in discovering if particular genetic attributes are associated with particular places. Such attributes develop over hundreds, if not thousands, of years through processes such as mutation and genetic drift (see Section 2.3), but their effects can be dampened by population mixing. Surnames provide enduring testimony to levels of mixing since their period of creation and can therefore be used to filter out the effects of several generations of migration. This section is concerned with the direct applications of the results and analysis outlined above and in previous chapters to the study of population genetics. Following a brief overview of the typical sample requirements for population genetics studies, this discussion will describe how individuals and areas can be targeted based on their surname distributions.

The use of surnames in population genetics has increased hugely (Colantonio *et al.* 2003). Such studies can say much about historical migrations and the degree to which different population groups have mixed on both national and international scales.

Until now little thought has been given to the value of spatial distributions of surnames in the context of sampling; something that appears to be related to a lack of georeferenced population data, or the ability to analyse it within population genetics. The sampling strategies of twelve studies (Lao *et al.* (2008), Novembre *et al.* (2008), Caravello and Tasso (1999), King and Jobling (2009), Lasker and Mascie-Taylor (1985), Sokal *et al.* (1992), Wilson *et al.* (2001), Rosser *et al.* (2000), Mascie-Taylor and Lasker (1990), Guglielmino and De Silvestri (1995), King and Jobling (2009), Capelli *et al.* (2003)) have been reviewed to establish common requirements in genetic sample design.

The most frequent sampling criterion is that an individual has ancestry, typically paternal grandparents, born within a specified distance, for example 20 miles, of their birthplace. This requirement is designed to screen out recent migrants to an area; if a person wishes to volunteer in a genetic study they will only be eligible to do so if both sets of grandparents were born close to where they too were born. This criterion is more likely to manifest distinctive regional genetic structure if both the person and their grandparents were born in an area where all of their surnames are most concentrated. If this is the case, it suggests that the person's ancestry has existed in that area since the surnames were first coined (see Winney *et al.* 2011).

Studies concerned with the links between surnames and the Y-Chromosome, outlined in more detail in Chapter 2.3, are focused upon males bearing surnames with a single origin (see Jobling (2001)). Previously, such information has only been available in a digital format through partial genealogical databases. An automated means to infer the origins of surnames from both contemporary and historical data sources therefore presents a valuable resource to such studies. Using the metrics outlined in Table 4-1 it is possible for the first time for researchers to select surnames according to a range of characteristics – of which single point of origin is one – in a matter of seconds. This process would hitherto have taken days of sifting through descriptions of surname characteristics from genealogical books or simple databases with no spatial referencing. On this basis it is now possible to select a large number of surnames and rank them in order of preference according to the desired metrics. Previous research has required a pre-conceived idea of the most appropriate

surnames for study, often based on anecdotal information to minimise the amount of background research.

In addition, the surname classifications outlined above provide a range of consistent indicators upon which to distinguish between rural and urban areas – a common requirement in population genetics studies. For example, Capelli *et al.* (2003) selected towns with populations of less than 20,000 and Lao *et al.* (2008) sampled from "rural" and "upland" populations. This reflects the fact that the populations of large urban areas are extremely diverse with many people being first generation migrants or with ancestry that originated elsewhere. Unfortunately the rural/ urban divide is ill-defined leading to inconsistent, and possibly incorrect classifications of a population as rural or urban. If the extent, or area of influence, for an urban area is perceived to be larger than is the case in reality, then the size of the population available for sampling will be unnecessarily small. Conversely, if an urban area's influence (in terms of population mixing) is smaller than anticipated it is likely that misleading samples will be collected.

The problem of defining the limits of an urban area is complex, but this thesis has provided two measures to help inform this. The first is linked to the insights into a surname's probable area of origin, provided by the KDE method, which enables researchers to isolate those surnames from an urban area that are native and ignore the surnames that have been imported from elsewhere. In addition, the results above suggest that surnames that occur in Great Britain as a result of international migration still exhibit spatial patterning. For example, their influence rarely extends beyond urban areas in the early years following migration, and they are amongst the most tightly clustered of all surnames. These patterns can be easily identified and accounted for. The second aspect of defining urban limits is related to the surname regions. In the case of Great Britain, as has already been discussed, urban areas have been clustered into their own distinct groupings consistent with the common surname compositions associated with them. Such limits may not follow administrative urban boundaries but nonetheless provide a useful definition, in the context of population mixing, of the different characteristics of urban areas. The classifications produced in this thesis can therefore offer researchers more

confidence in positioning rural/ urban boundaries, in addition to identifying diagnostic surnames that can be excluded from sampling because of the high likelihood that they, and therefore their (male) bearers, originated elsewhere.

As has been outlined in the previous section, the classification of urban areas forms only one aspect of the regional classifications produced: it also identifies areas relatively unaffected by migration. This provides an extra dimension to the cultural, and in the case of Europe, linguistic groupings confirmed by the surname regions produced in Chapters 5 and 6. The populations of such areas might be prioritised in population genetics studies as they are likely to have been relatively unaffected by external influences over time.

The methodological approaches utilised in this thesis therefore offer a large amount of previously unobtainable information that can be used to improve sampling in population genetics. By quantitatively establishing the points of origin, and the past and present extents of surnames in Great Britain, a hierarchy can be produced that identifies surnames whose bearers are likely to best represent the genetic composition of a particular population. In addition, the populations and areas they exist within can also be characterised by their likely interaction with surrounding groups. Information concerning the people to be sampled and the surname regions where they live has not previously been readily available and facilitates potentially impressive efficiency savings in future sample design of costly genetic sampling. Preliminary use of surnames, based on the results from this thesis, has already been successful in population genetics to the extent that Winney *et al.* (2011) state that:

"We believe that our method [based on surnames] *of selecting volunteers is a powerful way to collect a set of samples that can be used for high quality analysis of fine scale population structure in the UK. Subsequent localisation using surnames can sharpen the results of the* [genetic] *structure analysis."* (Winney *et al.* 2011: 15)

The above study applied surname a criterion to volunteers after they had been selected and sampled. An obvious efficiency saving, therefore, would be surname-based targeting of volunteers before they are sampled. This would result in the

application of similar strategies to those already used by private companies and public sector organisations that target provision through the use of geodemographic classifications. This would lead to a more efficient deployment of resources to the particular areas of interest in order, for example, to promote the study to potential volunteers.

In addition to targeting, the utility of both regional and individual representations of surnames also provides a means of better informing the selection of sample sites within any spatial sampling frame. Previous studies, such as Capelli *et al.* (2003), based their sampling strategy on an arbitrary grid common to ecological studies. As Figure 7-6 demonstrates, Capelli *et al.* (2003) sampled "small towns" that fell near to intersections on their grid. In the context of the findings of this thesis, especially in relation to regional geographies and the definitions of urban areas, it is clear that the sample locations shown in Figure 7-6 are almost entirely arbitrary, given the rich contextual information provided by surnames about the differences in population structure at a variety of scales, and the rate of both spatial and temporal changes.

Instead, based on the likely mobility of the populations contained within, particular areas can be selected in a systematic way to rule out the "noise" associated with widespread migrations. This is yet to be implemented in population genetics but would serve to create a sample design sensitive to both local variation and the global trends within the data. Such an approach can only serve to improve on the quality of the outcomes from current research.

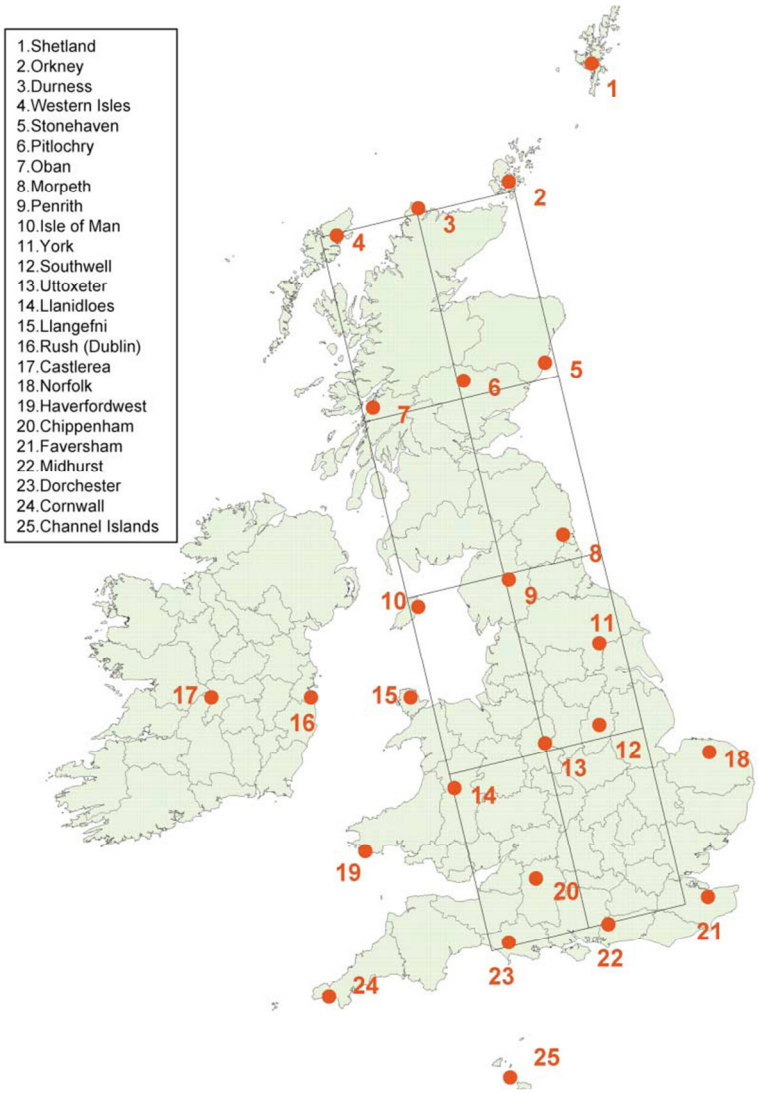


Figure 7-6: "*British Isles Sampling Locations Map: The location of the sampled small, urban areas and the 3 X 5 grid of collection points are shown. For each grid point, we selected the closest town within a 20-mile radius. Only towns with 5–20,000 inhabitants were chosen. Individuals were, with the exception of one location, then selected if their paternal grandfather's birthplace was within a 20-mile radius of the selected center. Midhurst samples were collected up to 40 miles from the respective grid point. When the grid point was at sea, the nearest point on the coast was used (Morpeth and Stonehaven). We also added additional points to cover important geographic regions not covered by the grid (Shetland, York, Norfolk, Haverfordwest, Llangefni, Chippenham, Cornwall, Channel Islands) and included two Irish samples, Castlerea and Rush (North of Dublin). The total number of points sampled in the British Isles was 25.*" Quoted from Capelli et al. (2003: 980).

7.2.3 TOWARDS A NATURAL UNIT OF ANALYSIS

Demonstrated here is the cultural, linguistic and genetic significance of surnames alongside a variety of methods able to unearth patterns in their spatial distributions. This renders them a plausible basis for defining "natural" units of analysis for studies into population genetics and studies of migration. What follows is a brief justification for this – perhaps bold – claim. It will first consider a number of possible extensions that may be required before surnames can be widely used to define valid spatial units before offering two examples: the first concerning migration and the second population genetics.

The creation of spatial units from the surname regions outlined in Chapters 5 and 6 is relatively straightforward but is subject to practical considerations. The first is the creation of aggregations sufficiently small to be of use in more detailed studies, and the second relates to contiguity constraints. In the context of European surname regions, these can be used as the most aggregate geographic units, equivalent to a country, and provide the basis for further clustering within each region to produce a nested set of spatial units. The number of these units would depend both on the application, but also how meaningful they are in the context of the surnames they are based on. Too many regions would result in noise and increasingly uncertain results as the populations they represent decrease in size and the basis on which to judge similarity or differences becomes increasingly susceptible to the problems associated with small numbers. In addition, contiguity constraints may be required to reduce the fragmentation of the spatial units at finer levels of granularity. Many of the findings from the creation of surname regions suggest that contiguity constraints should be used as a last resort and, therefore, they are only appropriate for specific applications. In many cases it would be acceptable to simply treat sets of spatial units that are not contiguous, but assigned to the same cluster, as different features.

In the context of geographically extensive studies of migration, it is conceivable that groups produced by comparative measures of surname compositions provide a strong indication of the degree of interaction between regions. In the case of Europe, spatial units assigned to different regions in Figure 6-7 could be assigned weights in

migration modelling on the assumption that interregional migration is less likely than intraregional migration. In this case Lasker Distance could be used as a substitute for geographic distance such that spatial units with more similar surname compositions might be treated as more similar entities. This would be especially useful in Central Europe, for example, where the rates of cultural and linguistic mixing are dependent on geopolitical factors as well as proximity.

Geographers have taken more interest in studies of migration using data other than surnames and have therefore created multiple regional geographies based on the principles of uniformity and function outlined in Section 5.3. Within population genetics no definitive regions have been created at the sub-national level and their utility not yet considered. The surname regions mapped previously in this thesis thus provide an important contribution and are a pragmatic "natural" unit around which to base genetic investigations. The power of such regionalisations lies in the fact that they represent a ubiquitous cultural attribute with demonstrable links to genetic characteristics. As Figure 7-7 demonstrates, the levels of aggregation provided by the regions mapped in Chapters 5 and 6 are likely to be appropriate. This is because geographic variations in genetic structure are relatively small and therefore require samples to be drawn from a large area in order for the differences to be readily measurable. It is therefore both appropriate and straightforward to integrate the results of this thesis into future population genetics research.

7.3 FUTURE WORK

This thesis has highlighted the value of surnames as a spatial data source, whilst seeking to address many of the shortcomings of previous research. Despite the unprecedented detail of the analysis provided, it has nevertheless only been possible to provide partial insights into the information that can be obtained from the spatial distributions of surnames. The final section of this chapter will outline a number of avenues for further research, in terms of both the methods employed to investigate surnames and the application of the results.

7.3.1 METHODS AND DATA

A range of methods have been utilised that successfully demonstrate both the spatial clustering of individual surnames and regional similarities in surname compositions. Future work would therefore be most productive by extending current methods rather than devising more novel approaches.

A natural extension to the methodological approaches outlined above would be their combined use to enhance some of the patterns shown. Surnames with multiple origins, for example, obscure the contrasts between the pairwise distances contained in the Lasker Distance matrix (Manni *et al.* 2005). Manni *et al.* (2005) argue that, from a population genetics perspective, the distinctions between the resulting regions are reduced because surnames with multiple origins give rise to an artificial kinship between different populations. With the KDE method outlined in Section 4.1.5, however, it would be straightforward to identify such surnames for the purposes of their exclusion from the Lasker Distance calculation. On this basis an initial filtering step using the KDE method can be applied to remove surnames with multiple cores before undertaking clustering.

The (dis)similarity between the surname compositions of populations has been established between areas with the Lasker Distance (Equation 5.2). It was acknowledged that alternative measures, such as the Nei's Distance, are available and

the effect of their use in classification should be explored in further work. In addition, a more complex issue, alluded to in the discussion of the European data (Section 6.2.1), is the handling of the different levels of aggregation associated with the inconsistent population sizes represented by each spatial unit. Dissimilarity measures, such as the Lasker Distance, rely on comparisons between aggregate population groups that are often equally weighted for the analysis. A spatial unit representing 100 people is therefore treated in the same way as one with 1,000 or even 10,000. As was discussed in Chapter 6.2, in the European case a country's influence on the analysis is in part based on the number of spatial units it has rather than the size of its population. The likely result is an apparent increase in diversity for countries partitioned into large numbers of regions, despite relatively uniform surname compositions.

A number of approaches could be used to mitigate the drawbacks associated with inconsistent levels of aggregation within distance measures. The obvious solution would be the greater standardisation of spatial units across Europe, in order that they better reflect population density. This, however, leads to complications such as whether the size of the resulting units should reflect the target population density or the sampled population density. In addition, more sparsely populated areas are going to require larger units (in terms of geographic extent) in order to meet a population threshold and this is likely to risk amalgamating culturally distinct groups as potential surname boundaries are crossed. This solution would present a major undertaking at the European level and may not produce significantly improved results. More practical options could therefore include weighting the dissimilarity calculation or its subsequent clustering. One possible approach, in this context, would simply be to multiply the elements of the Lasker Distance matrix by a suitably normalised population weight. Such an approach may also require some nationally varying "alpha" value to alter the influence of the population weighting on the cluster outcome.

In addition to the more complex issues outlined above, there are a number of straightforward improvements that can be made to develop the work in this thesis. The first would be the creation of a complete European regionalisation by including

countries such as Portugal and Finland. The methodological steps to geocode the data and, as a result of this thesis, analyse it are established; the challenge lies in sourcing and purchasing such commercially and often personally sensitive data. A related advancement of the results presented in this thesis would be to pursue more localised studies, perhaps along the boundaries of the large-scale clusters, to establish their strength at finer levels of granularity. This has already been discussed to a limited extent in the context of transitions in urban areas, but those abutting rural features, such as the Danelaw line in Figure 6-3, would benefit from further research. Such studies may not warrant different methods but would require the use of less coarse data, especially in the European context, and could be placed in the framework of other data sources such as digital elevation models to represent topographic boundaries.

7.3.2 APPLICATIONS

Whilst there is inevitably room for methodological refinement to the approaches taken in this thesis, the insights already offered into the spatial behaviour of surnames present a firm basis for future research. One of the listed aims of this research (see Chapter 1) is to provide stimulus for hypothesis generation to facilitate future research incorporating surnames. Much of the content has concerned the theoretical and methodological approaches required to produce robust indicators of both the individual and the regional geographies associated with surnames. Applications of the results have not been covered in the same depth and provide much to focus on in the future.

One of the applications outlined in most depth is the potential for the insights provided by surnames to be applied to population genetics, especially with regard to the recruitment and sampling of volunteers. The extent to which surnames enhance the quality of the results in such studies has only been subject to preliminary analysis (see Winney *et al.* 2011) and therefore requires further research. Additionally, as the quality and volume of genetic samples increase it should be possible to begin to explore the relationship between surname regions and their genetic counterparts. The

genetic differences will be far less profound but may provide interesting insights into population movements that pre-date the creation of surnames.

The use of surnames to infer patterns of migration has been discussed in detail above and in Section 4.3. In addition, a number of other studies reviewed in Darlu *et al.* (2011), have used a modelling approach using Bayesian statistics and logistic regression to infer historic migration using surnames. Such methods rely on data being recorded at two time periods and focus on inferring movements from areas where the surname was present in the first time period to where they were absent in the first time period but present in the second time period. The use of both surname regions and their core concentrations can supplement the simple counts that underlie the methods reviewed by Darlu *et al.* (2011) by providing regions of increased interaction between surnames, something they assume to conform to arbitrary administrative units. In addition, the probable surname origins produced with KDE can be used to extend the analysis back further than the oldest dataset available in the study. Future work could also investigate the possibility of improving the prediction of migration flows based on surname regions as natural units of analysis, described in Section 4.3.

A final logical extension to this work is the exploration of the contemporary significance of surnames in the context of geodemographics. Such research would expand on that of Longley *et al.* (2007) who charted contemporary prospects for descendants of the migrants who moved from Cornwall to Middlesbrough. On the basis that urban areas do not present a homogenous distribution of surnames with many bearers continuing to cluster near those with similar cultural attributes it follows that the impact of migration in terms of lifestyle choices and life chances will have differing effects on different groups. Such groups can be easily identified and classified according to the characteristics of their surnames. In the case of the Scottish migration to Corby it would, for example, be interesting to see if raised levels of unemployment exist within a community disproportionately affected by the second round of steelwork closures. Corby represents an obvious example but there may be many subtler movements highlighted by the shifts in surname cores (see Figure 4-26) that indicate movement due to the specific economic circumstances.

The suggested avenues of future research outlined above are by no means the only opportunities to develop the research presented in this thesis. The aim here has been to introduce a number of ways in which the analysis of surname distributions can better inform population research. The insights, both technical and conceptual, outlined in the first part of this chapter enforce the methodological contributions of this thesis both in the field of surnames research and geography more broadly. Such insights, when combined with the contemporary relevance of surnames in a variety of fields, such as population genetics and studies of migration, confirm the importance of surnames as a source of quantitative spatial data.

8 THESIS SUMMARY AND CONCLUSIONS

The spatial analysis of surname distributions presented in this thesis is unprecedented in its detail and extent. The results characterise more people and places according to the geographies of their surnames than ever before. This final chapter draws together the multiple research strands of this thesis to assess the extent to which they address the four aims stated in the introduction. Each aim is re-stated and discussed before final conclusions are offered.

1. *To review previous surnames research in the context of spatial analysis and quantitative geography more broadly.*

The aim above was perhaps the most straightforward to fulfil because very little surname research has been framed in the context of quantitative geography. As Chapter 2 demonstrates, this is surprising because quantitative geography in the past has been accused of being methods rich but data poor and surname frequencies offer a comprehensive source of data well suited to such methods. The topics covered in this thesis are therefore timely through their application, and development, of largely established spatial analysis methods to hundreds of millions of individuals based on surnames. Promoting the synergy between spatial analysis and surnames also marks an important contribution to surnames research by affirming the inherently spatial nature of surnames, despite many of the past studies of their spatial behaviour appearing tenuous in the context of mainstream geographical analysis. Such limitations, however, fail to detract from the importance of surnames as cultural and genetic markers and do not conceal their potential utility as a source of spatial data worthy of further investigation.

2. *To create a methodology for the automated identification of key characteristics pertaining to individual surnames, such as extent and area(s) of highest concentration.*

Chapter 4 addressed this aim by first considering the range of novel (in the context of surname analysis) methods, such as the location quotient (LQ) and the detection of surface discontinuities, designed to identify spatial clusters. Of the selection tested, kernel density estimation (KDE) was considered the most appropriate and has been successfully integrated into a methodology capable of extracting key characteristics of individual surname distributions. The resulting methodology is robust, replicable and entirely automated. This latter point is significant as no previous methods have been able to wholly remove the need for manual disaggregation of surname distributions- an important consideration given the volume of surnames (nearly 50,000 across both 1881 and 2001). Identification of the core areas of concentration for surnames in 1881 and 2001 serves to demonstrate the relatively static geographies of many surnames, and also the nature of migration. The digital storage of the results enables them to be easily queried by future researchers to facilitate their use in a range of applications.

3. *To review and establish the methodological processes required for the aggregation and regionalisation of surnames appropriate to a range of scales and data sources.*

This was given the most extensive consideration in this thesis through Chapters 5 and 6. The work represents a direct advancement of past surnames research by replicating many of the results from its disparate methods with a single, cohesive, methodological framework. The results are unprecedented in terms of their extent and granularity, in addition to the relative completeness of the datasets used. This marked improvement on previous attempts at regionalising surnames increases confidence in the utility of the results in a range of applications, not least in comparison with historical population structure, assessing the impact of migration and providing links to the likely cultural similarities or differences between areas.

4. *To promote the outcomes from aims 2 and 3 as a basis for future research and hypothesis generation relevant to both a range of applications and disciplines.*

In many ways surname research is still in its infancy and this thesis only takes the initial steps required for surnames to be a widely utilised source of quantitative spatial data. This is largely reflected by the methodological focus at the expense of in depth treatment of applications. Where applications have been demonstrated, especially in the context of population genetics and the investigation of population continuity and change, the utility of surnames becomes clear. It is hoped therefore that Chapter 7 illustrates how the results of the previous chapters can provide firm foundations for future research beyond this thesis. Applications and further research include gauging the effects of migration through assessing levels of population continuity and change and also providing a basis on which to develop an efficient sampling design for studies in population genetics. Perhaps one of the most interesting applications of the results is the use of surname regions as natural units of analysis in population studies. The cultural, linguistic and genetic significance of surnames makes them a more meaningful metric on which to study population than often-arbitrary administrative boundaries.

8.1 FINAL CONCLUSIONS

This thesis is evidence for the depth of insight that can be gained from the appropriate spatial analysis of a previously overlooked data source comprising surnames, their locations and their frequency. Based on the persistence of the patterns over time, in the case of Great Britain, and the broad conformity to language and culture (sometimes at odds with national borders), in the case of Europe, the results from this thesis suggest that surnames offer an indicator of historic population structure along with all the ephemera of subsequent economic and social geographies heaped upon it. Such a claim can be confidently based on the firm, replicable, analysis of an unprecedented volume of data.

In many ways it is remarkable that something as ubiquitous and seemingly innocuous as a surname can reveal information at the individual level concerning ancestral origins, migration history, and genetic relatedness to others. At different aggregate levels, surnames also offer insights into the short and long term dynamics of population mobility and mixing. Previously, such information has been partially obtained using small samples of surnames requiring many hours of manual processing. Demonstrated here are a series of inductive computational methodologies capable of discerning the same information applicable to complete population datasets (as opposed to samples), created within just a few seconds of processing for each surname or unit of population. The results differ little from the indications provided by more partial research and present the most comprehensive picture of Great British and European surname geography to date.

In conclusion, it is clear that surnames should no longer be overlooked by quantitative geography a means to discern population characteristics. This thesis has unearthed an enduring population structure, applicable to a range of spatial and temporal scales. This is an achievement that provides firm foundations for future research and applications in a range of disciplines.

9 REFERENCES[4]

Adnan, M. 2011. *Towards Specification, Design, and Testing of Real-Time Geodemographics*. UCL Geography PhD Thesis (In Prep.).

Adnan, M., Longley, P., Singleton, A. and Brunsdon, C. 2010. Towards Real-Time Geodemographics: Clustering Algorithm Performance for Large Multidimensional Spatial Databases. *Transactions in GIS*. 14: 283-297.

Bação F, Lobo V, Painho M. 2004. Clustering Census Data: Comparing the Performance of Self-Organising Maps and *K*-means Algorithms. *KDNet Symposium*, Bonn, Germany.

Bação F, Lobo V, Painho M. 2005. Self-Organizing Maps as Substitutes for *K*-means Clustering. *Lecture Notes in Computer Science* 3416: 476-483.

Barbujani, G. and Sokal, R. 1990. Zones of Sharp Genetic Change in Europe are also Linguistic Boundaries. *Proceedings of the National Academy of Science*. 87: 1816-1819.

Barker, S., Spoerlein, S., Vetter, T., Viereck, W. 2007. *An Atlas of English Surnames*. Peter Lang: Frankfurt.

Barrai, I., Rodriguez-Larralde, A., Mamolini, E., Manni, F., Scapoli, C. 2000. Elements of the Surname Structure of Austria. *Annals of Human Biology*. 27, 6: 607-622.

Barrai, I., Rodriguez-Larralde, A., Manni, F., Scapoli, C. 2002. Isonymy and Isolation by Distance in the Netherlands. *Human Biology*. 74, 2: 263-283.

[4] Papers with a large number of authors are listed in the format "First Author" *et al.*

Barrai, I., Rodriguez-Larralde, A., Manni, F., Ruggerio, V., Tartari, D., Scapoli, C. 2004. .Isolation by Language and Distance in Belgium. *Annals of Human Genetics.* 68, 1: 1-16.

Bentley, A., Ormerod, P. and Batty, M. 2011. Evolving Social Influence in Large Populations. *Behavioural Ecology and Sociobiology.* 65: 537-546.

Bivand, R., Pebesma, E. and Gomez-Rubio, V. 2008. *Applied spatial data analysis with R.* Springer: New York.

Bowden, G. *et al.* 2008. Excavating Past Population Structures by Surname-Based Sampling: The Genetic Leagcy of the Vikings in Northwest England. *Molecular Biology and Evolution.* 25, 2: 301-309.

Bowman, A. and Azzalini, A., 1997. *Applied Smoothing Techniques for Data Analysis: the Kernel Approach with S-Plus Illustrations*, Oxford University Press: Oxford

Bracken, I. and Martin, D. 1995. Linkage of the 1981 and 1991 UK Censuses Using Surface Modelling Concepts. *Environment and Planning A.* 27: 379-390.

Branco, C. and Mota-Vieira, L. 2004. Population Structure of São Miguel Island, Azores: A Surname Study. *Human Biology* 75, 6: 929-939.

Brassel, K. and Reif, D. 1979. A Procedure to Generate Thiessen Polygons. *Geographical Analysis.* 11, 3: 289-303.

Brown, L. and Holmes, J. 1971. The Delimination of Functional Regions, Nodal Regions, and Hierarchies by Functional Distance Approaches. *Journal of Regional Science.* 11, 1: 57-72.

Brunsdon, C. 2009. Finding Fault: Identifying and Testing 'Social Faultiness' in Surface Fitting Techniques. In Lees, B.G. and Laffan, S.W. (eds). *10th International Conference on GeoComputation*, UNSW, Sydney, November-December 2009.

Burt, J., Barber, G. and Rigby, D. 2009. *Elementary Statistics for Geographers* (3rd ed.). Guilford Press: London.

Capelli *et al.* 2003. A Y Chromosome Census of the British Isles. *Current Biology*. 13: 979-984.

Caravello, G. and Tasso, M. 1999. An analysis of the spatial distribution of surnames in the Lecco area (Lombardy, Italy). *American Journal of Human Biology*, 11, 3: 305-315.

Cavalli-Sforza, L. 2000. *Genes, Peoples and Languages*. Penguin Books, London.

Cavalli-Sforza, L., Feldmen, M, Chen, K. and Dornsbusch, S. 1982. Theory and Observation in Cultural Transmission. *Science.* 218, 19-27.

Chainey, S., Tompson, L. and Uhlig, S. 2008. The Utility of Hotspot Mapping for Predicting Spatial Patterns of Crime. *Security Journal*, 21, 1-2: 4-28.

Cheshire, J., Adnan, M. and Gale, C. 2011. The Use of Consensus Clustering in Geodemographics. *Proceedings of GIS Research UK 2011, University of Portsmouth.* 45-62.

Cheshire, J. and Longley, P. 2011. Identifying Spatial Concentrations of Surnames. *International Journal of GIS* (In Press).

Cheshire, J., Singleton, A. and Longley, P. 2010. The Surname Regions of Great Britain. *Journal of Maps*. Doi 10.4113/jom.2010.1103.

Clauset, A., Shalizi, C. and Newman, M. 2009. Power-Law Distributions in Empirical Data. *SIAM Review* 51, 4: 661-703.

Claval, P. 1998. *An Introduction to Regional Geography*. Blackwell: Oxford.

Cliff, A. and Ord, J. 1981. *Spatial Processes.* Pion: London.

Cloke, P., Philo, C., Sadler, D. 1991. *Approaching Human Geography.* Sage, London.

Colantonio, S., Lasker, G., Kaplan, B., Fuster, V. 2003. Use of Surname Models in Human Population Biology: A Review of Recent Developments. *Human Biology.* 75, 6: 785-787.

Coleman, D. and Haskey, J. 1986. Marital Distance and its Geographical Orientation in England and Wales, 1979. *Transactions of the Institute of British Geographers.* 11, 3: 337-355.

Coleman, D. and Salt, J. 1992. *The British Population: Patterns, Trends and Processes.* Oxford University Press: Oxford.

Crow, J., Mange, P. 1965. Measurement of Inbreeding from the Frequency of Marriages Between Persons of the Same Surname. *Eugenics Quaterly.* 12:199-203.

Darby, H. 1973. *A New Historical Geography of England.* Cambridge University Press, Cambridge.

Darlu, P., Brunet, G., Barbero, D. 2011. Spatial and Temporal Analyses of Surname Distributions to Estimate Mobility and Changes in Historical Demography: The Example of Savoy (France) from the Eighteenth to the Twentieth Century. In Gutmann *et al.* (eds.). *Navigating Time and Space in Population Studies,* International Studies in Population 9. Springer, New York.

Darwin, G. 1875. Marriages Between First Cousins in England and Their Effects. *Journal of the Statistical Society.* 38, 2: 153-184.

de Smith, M., Longley, P and Goodchild, M. 2009. *Geospatial analysis: a comprehensive guide to principles, techniques and software tools* (2nd ed.), Leicester: Matador.

Dorling, D. 1995. *A New Social Atlas of Britain.* Wiley: Chichester.

Everitt, B. 1972. Cluster Analysis: A Brief Discussion of Some of the Problems. *British Journal of Psychiatry.* 120: 143-145.

Everitt, B., Landau, S., Leese, M. 2001. *Cluster Analysis* (4th ed.). Hodder, London.

Fotheringham, S. 2006. Quantification, Evidence, Positivism. In Aitken, S. and Valentine, G. (eds.) 2006. *Approaches to Human Geography.* 237. Sage: London.

Fotheringham, S., Brunsdon, C., Charlton, M. 2007. *Quantitative Geography Perspectives on Spatial Data Analysis.* Sage, London.

Fox, W. and Lasker, G. 1983. The Distribution of Surname Frequencies. *International Statistical Review.* 51: 81-87.

Fryer, Roland G. and Levitt, S. 2004. The Causes and Consequences of Distinctively Black Names. *The Quarterly Journal of Economics* .119, 3: 767.

Gastner, M. and Newman, M. 2004. From The Cover: Diffusion-based method for producing density-equalizing maps. *Proceedings of the National Academy of Sciences*, 101(20), 7499-7504.

Gatrell, A. C. 1981. Multidimensional Scaling. In Wrigley, N., and Bennett, R. (eds), *Quantitative Geography: a British View.* 151. Routledge and Kegan Paul, London.

Gilbert, A. 1988. The New Regional Geography in English and French Speaking Countries. *Progress in Human Geography.* 12, 208: 208-228.

Golledge, R., Rushton G. 1972. Multidimensional Scaling: Review and Geographical Applications. *Association of American Geographers Commission on College Geography, Technical Paper No. 10.*

Goodchild, M. 1986. Spatial Autocorrelation. *CATMOG* 47. GeoBooks: Norwich.

Goodchild, M., Anselin, L. and Deichmann, U. 1993. A Framework for the Areal Interpolation of Socioeconomic Data. *Environment and Planning B.* 25: 383-397.

Goodchild, M., Yuan, M. and Cova, T. 2007. Towards a General Theory of Geographic Representation in GIS. *International Journal of Geographical Information Science*, 21, 3: 239-260.

Gordon, A. 1987. A Review of Hierarchical Classification. *Journal of the Royal Statistical Society. Series A (General).* 150, 2: 119-137.

Gordon, A. 1999. *Classification (2nd Edition).* Chapman and Hall, London.

Grieco, M. 1985. Corby: New Town Planning and Imbalanced Development. *Regional Studies.* 1: 9-18.

Grigg, D. 1965. Regions, Models and Classes. In Chorely, R., Haggett., P. (eds). 1965. *Models in Geography.* Methuen and Co: London.

Grigg, D. 1967. The Logic of Regional Systems. *Annals of the Association of American Geographers.* 55, 3: 465-491.

Guglielmino, C. and De Silvestri, A. 1995. Surname sampling for the study of the genetic structure of an Italian province. *Human Biology; an International Record of Research.* 67, 4: 613-628.

Guppy, H., 1890. *Homes of Family Names in Britain.* Harrison and Sons: London.

Haggett, P. 1965. *Locational Analysis in Human Geography* (1st ed.). Arnold: London.

Haggett, P. 1994. Prediction and Predictability in Geographical Systems. *Transactions of the Institute of British Geographers.* 19, 1: 6-20.

Haggett, P., Cliff, A. and Frey, A. 1977. *Locational Analysis in Geography 2: Locational Methods*. Edward Arnold: London.

Haining, R. 2009. Special Nature of Spatial Data. In Fotheringham, S. and Rogerson, P. (eds.). *The Sage Handbook of Spatial Analysis*. 5. Sage: London.

Harris, R., Sleight. P., Webber, R. 2005. *Geodemographics, GIS and Neighbourhood Targeting*. John Wiley and Sons: Chichester.

Hartigan, J. A. and Wong, M. A. 1979. A *K*-means clustering algorithm. *Applied Statistics* 28**:** 100–108.

Hey, D. 2000. *Family Names and Family History*. Hambledon Continuum: London.

Hill, E., Jobling, M. and Bradley, D. 2000. Y-Chromosome Variation and Irish Origins. *Nature* 404: 351.

Jobling, M. 2001. In the Name of the Father: Surnames and Genetics. *Trends in Genetics*. 17, 6: 353-357

Johnston, R. 1968. Choice in Classification: The Subjectivity of Objective Methods. *Annals of the Association of American Geographers*. 58, 3: 579-589.

Johnston, R. 1970. Grouping and Regionalizing: Some Methodological and Technical Observations. *Economic Geography*. 46: 293-305.

Johnston, R., Gregory, D., Pratt, G. and Watts, M. 2005. *The Dictionary of Human Geography*. Blackwell; Oxford.

Jombart, T. 2008. adegenet: A R Package for the Multivariate Analysis of Genetic Markers. *Bioinformatics* 24: 1403-1405.

Kaplan, B. and Lasker, G. 1983. The Present Distribution of Some English Surnames Derived From Place Names. *Human Biology* 55, 2: 243-250.

Kaufman , L. and Rousseeuw, P. 1990. *Finding Groups in Data*. Wiley: New York.

Kelsall, J. and Diggle, P. 1995. Non-Parametric Estimation of Spatial Variation of Relative Risk. *Statistics in Medicine*, 14, 21-22: 2335-2352.

Keynes, S. 1997. The Vikings in England c.790-1016. In Sawyer, P.(ed). *The Oxford Illustrated History of the Vikings*. 48-82.

King, T. and Jobling, M. 2009. Founders, Drift and Infidelity: The Relationship Between Y Chromosome Diversity and Patrilineal Surnames. *Molecular Biology and Evolution*. 26, 5: 1093-1102.

Kleiweg, P., Nerbonne, J. and Bosveld, L. 2004. Geographic Projection of Cluster Composites. In Blackwell, A., Marriott, K., Shimojima, A. (eds) *Diagrams 2004, Lecture Notes in Computer Science*. Springer, New York.

Krygier, J. and Wood, D. 2011. *Making Maps* (2nd Edition). Guilford Press: New York.

Lankford, P. 1969. Regionalisation: Theory and Alternative Algorithms. *Geographical Analysis*. 1: 196–212.

Lao, O. *et al.* 2008. Correlation between Genetic and Geographic Structure in Europe. *Current Biology*, 18,16: 1241-1248.

Lasker, G. 1968. The Occurrence of Identical (Isonymous) Surnames in Various Relationships in Pedigrees: A Preliminary Analysis of the Relation of Surname Combinations to Inbreeding. *American Journal of Human Genetics*. 20:250–257.

Lasker, G. 1985. *Surnames and Genetic Structure.* Cambridge University Press, Cambridge.

Lasker, G. 1999. The Hierarchical Structure of an Urban Town, Kidlington, Oxfordshire Examined by the Coefficient of Relationship by Isonymy. *Journal of Biosocial Science.* 31: 279-284.

Lasker, G., 2002. Using Surnames to Analyse Population Structure. In Postle, D. (ed) *Naming, Society and Regional Identity.* Leopard's Head Press: Oxford.

Lasker, G and Mascie-Taylor, C., 1985. The Geographical Distribution of Selected Surnames in Britain. Model Gene Frequency Clines. *Journal of Human Evolution.* 14: 385-292.

Lauderdale, D. S. and Kestenbaum, B. 2000. Asian American ethnic identification by surname. *Population Research and Policy Review* 19, 3: 283- 291.

Lloyd, C. 2007. *Local models for spatial analysis.* CRC/Taylor and Francis. London.

Longley, P., Cheshire, J. and Mateos, P. 2011a. Creating a Regional Geography of Britain Through the Spatial Analysis of Surnames. *Geoforum.* (In Press). Doi 10.1016/j.geoforum.2011.02.001.

Longley, P., Goodchild, M., Maguire, D. and Rhind, D. 2011b. *Geographic Information Systems and Science* (3rd ed.). Wiley: New York

Longley, P., Webber, R., Lloyd, D. 2007. The Quantitative Analysis of Family Names: Historic Migration and the Present Day Neighborhood Structure of Middlesbrough, United Kingdom. *Annals of the Association of American Geographers.* 97, 1: 31-48.

Lower, M. 1860. *Patronymica Britannica.* John Russell Smith: London.

Lu, H. and Carlin, P. 2005. Bayesian Areal Wombling for Boundary Analysis. *Geographical Analysis.* 37: 265-285.

Macintyre, S. and Sooman, A. 1991. Non-Paternity and Prenatal Genetic Screening. *The Lancet.* 338, 8771: 869-871.

MacLeod, G. and Jones, M. 2001. Renewing the Geography of Regions. *Environment and Planning D: Society and Space.* 19: 669-695.

MacQueen J. 1967. Some Methods for classification and analysis of multivariate observations. *Proceedings from the 5th Berkeley Symposium on Mathematical Statistics and Probability.* 281-297.

Manel, S., Schwartz, M., Luikart, G., Taberlet, P. 2003. Landscape Genetics: Combining Landscape Ecology and Population Genetics. *Trends in Ecology and Evolution.* 18, 4: 189-197.

Manni, F. and Barrai., I. 2001. Genetic Structures and Linguistic Boundaries in Italy: A Microregional Approach. *Human Biology.* 73, 3: 335-347.

Manni, F., Guerard, E. and Heyer, E. 2004. Geographic Patterns of (Genetic, Morphologic, Linguistic) Variation: How Barriers Can Be Detected by Using Monmonier's Algorithm. *Human Biology.* 76, 2: 173-190.

Manni, F., Heeringa, W. and Nerbonne, J. 2006. To What Extent are Surnames Words? Comparing Geographic Patterns of Surname and Dialect Variation in the Netherlands. *Literary and Linguistic Computing.* 21, 4: 507-528.

Manni, F. Heeringa, W. Toupance, B. Nerbonne, J. 2008. Do Surname Differences Mirror Dialect Variation? *Human Biology* 80, 1: 41-64.

Manni, F., Toupance, B., Sabbagh, A., and Heyer, E. 2005. New Method for Surname Studies of Ancient Patrilineal Population Structures, and Possible

Application to Improvement of Y-Chromosome Sampling. *American Journal of Physical Anthropology*. 126: 214- 228.

Manrubia, S. and Zannette, D. 2002. At the Boundary Between Biological and Cultural Evolution: The Origin of Surname Distributions. *Journal of Theoretical Biology*. 216, 461-477.

Martin, D. 1996. *Geographic Information Systems: Socioeconomic Applications* (Second Edition). Routledge: New York.

Martin, D. 1998a. Automatic neighbourhood identification from population surfaces *Computers, Environment and Urban Systems* 22, 107-120.

Martin, D. 1998b. Optimising Census Geography: the Separation of Collection and Output Geographies. *International Journal of Geographic Information Science*. 12, 7: 673-685.

Martin, D. 1999. Spatial representation: the social scientist's perspective In: Longley, P., Goodchild, M., Maguire, D. and Rhind, D. (eds.) *Geographical Information Systems: Principles, Techniques, Applications and Management* (2nd ed). 71. Wiley: Chichester.

Martin, D. 2002. Geography for the 2001 Census in England and Wales. *Population Trends*. 108: 7-15.

Mascie-Taylor, C., Boyce, A., Brush, G. in Lasker, G. 1985. *Surnames and Genetic Structure*. Cambridge University Press, Cambridge.

Mascie-Taylor, C. and Lasker, G. 1990. The Distribution of Surnames in England and Wales: A Model for Genetic Distribution. *Man*, 25, 521-530.

Massey, D. 1995. *Spatial Divisions of Labour, Social Structures and the Geography of Production* (2nd ed.). Palgrave Macmillan: London.

Massey, D. and Jess, P. (eds). 1995. *A Place in the World? Places, Cultures and Globalisation.* Oxford University Press: Oxford.

Mateos, P. And Tucker, D.K. 2008. Forenames and Surnames in Spain in 2004. Names, a Journal of Onomastics. 56, 3: 165-184.

McClure, P. 1979. Patterns of Migration in the Late Middle Ages: The Evidence of English Place-Name Surnames. *The Economic History Review.* 32, 2:167-182.

McElduff, F., Mateos, P., Wade, A., Cortina Borja, M. 2008 What's In a Name? The Frequency and Geographic Distributions of UK Surnames. *Significance.* 5, 4: 189-192

McQuitty, L. 1957. Elementary Linkage Analysis for Isolating Orthogonal and Oblique Types and Typal Relevancies. *Educational and Psychological Measurement.* 17:207-229.

Milligan, G. 1980. An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms. *Psychometrika.* 45: 325-342.

Monmonier, M. 1973. "Maximum-Difference Barriers: An Alternative Numerical Regionalisation Method." *Geographical Analysis* 5, 3: 245-261.

Monti, S., Tamayo, P., Mesirov, J., Golub, T. 2003. Consensus Clustering: A Resampling Based Method for Class Discovery and Visualisation of Gene Expression Microarray Data. *Machine Learning.* 52: 91-118.

Moore, L., McEvoy, B., Cape, E., Simms, K. and Bradley, D. 2006. A Y-Chromosome Signature of Hegemony in Gaelic Ireland. *American Journal of Human Genetics.* 334-338.

Moran, P. 1948. The Interpretation of Statistical Maps. *Journal of the Royal Statistical Society. Series A. General*, 10(2), 243.

Murphy, A. 1991. Regions as Social Constructs: the Gap Between Theory and Practice. *Progress in Human Geography*. 15, 1:22-35.

Nei, M., 1973. The Theory and Estimation of Genetic Distance. In *Genetic Structure of Populations*. Edited by Morton, N. E. 45-64.

Nei, M., 1978. Estimation of Average Heterozygosity and Genetic Distance from a Small Number of Individuals. *Genetics*. 583-590.

Nerbonne, J., Kleiweg, P., Heeringa, W. and Manni, F. 2008.Projecting Dialect Distances to Geography: Bootstrap Clustering vs. Noisy Clustering. *Data Analysis, Machine Learning and Applications*. 647-654.

Novembre, J., Johnson, T., Bryc, K., Kutlaik., Z., Boyko, A., Auton, A., Indap, A., King, K., Bergmann, S., Nelson, M., Stephens, M. and Bustamante, C. 2008. Genes Mirror Geography Within Europe. *Nature*. 456: 98-101.

Openshaw, S. 1984. The Modifiable Areal Unit Problem, *CATMOG* 38. GeoBooks: Norwich.

Panaretos, J. 1989. On the Evolution of Surnames. *International Statistical Review*. 57, 2: 161-167.

Peña, J., Lozan, J., Larrańaga, P. 1999. An Empirical Comparison of Four Initialisation Methods for the *K*-means Algorithm. *Pattern Recognition Letters*. 20, 10:1027-1040.

Perona, P. and Malik, J. 1990. Scale-Space and Edge Detection Using Anisotropic Diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 12, 7: 629-639.

Piazza, A., Rendine, S., Zei, G., Moroni, A., Cavalli-Sforza, L. 1987. Migration Rates of Human Populations from Surname Distributions. *Nature*. 329: 714-716.

Pocock, D. 1960. The Migration of Scottish Labour to Corby New Town. *Scottish Geographical Journal.* 76, 3: 169-171.

Pooley, C., Turnbull, J. 1998. *Migration and Mobility in Britain Since the 18th Century.* UCL Press: London.

Porteus, J. 1982. Surname Geography: a Study of the Mell Family Name c. 1538-1980. *Transactions of the Institute of British Geographers.* 7, 4: 395-418.

Pudup, M. 1988. Arguments Within Regional Geography. *Progress in Human Geography* 12: 369-390.

R Development Core Team 2011. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Rodriguez-Larralde, A., Barrai, I., Nesti, C. Mamolini, E. and Scapoli, C. 1998. Isonymy and Isolation By Distance in Germany. *Human Biology.* 70, 6: 1041-1056.

Rodriguez-Larralde, A., Gonzales-Martin, A., Scapoli, C., Barrai, I. 2003. The names of Spain: a study of the isonymy structure of Spain. *American Journal of Physical Anthropology.* 121: 280-292.

Rodriguez-Larralde, A., Pavesi, A., Siri, G., Barrai., I. 1994. Isonymy and the Genetic Structure of Sicily. *Journal of Biosocial Science.* 26: 9-24.

Rodriguez-Larralde, A., Scapoli, C., Beretta, M., Nesti, C., Mamolini, E., Barrai, I. 1998. Isonymy and the genetic structure of Switzerland. II. Isolation by distance. *Annals of Human Biology.* 6: 533-540.

Rogers, A. 1991. Doubts about Isonymy. *Human Biology.* 63, 5: 663-668.

Rogerson, P. 2006. *Statistical Methods in Geography* (2nd ed). Sage Publications: London.

Rogerson, P. and Yamada, I. 2009. *Statistical Detection and Surveillance of Geographic Clusters.* CRC Press: Boca Raton.

Rosser, Z. *et al.* 2000. Y-Chromosonal Diversity in Europe Is Clinal and Influenced Primarily By Geography, Rather than by Language. *American Journal of Human Genetics.* 67: 1526-1543.

Scapoli, C., Goebl, H., Mamolini, E., Rodriguez-Larralde, A., Barrai, I. 2005. Surnames and Dialects in France: Population Structure and Cultural Evolution. *Journal of Theoretical Biology.* 237, 2: 75-86.

Scapoli, C., Mamolini, E., Carrieri, A., Rodriguez-Larralde, A., Barrai, I. 2007. Surnames in Western Europe: A Comparison of the Subcontinental Populations through Isonymy. *Theoretical Population Biology* 71, 37-48.

Scott, D., 1992. *Multivariate Density Estimation: Theory, Practice, and Visualisation*, Wiley: New York.

Shimodaira, H. 2004. Approximately Unbiased Tests for Regions Using Multiscale Bootstrap Resampling. *Annals of Statistics.* 32, 2616-2641.

Simpson, I., Armstrong, D., Jarman, A. 2010. Merged Consensus Clustering to Assess and Improve Class Discovery with Microarray Data. *BMC Bioinformatics*, 11: 590.

Singleton, A., Longley, P. 2008. Creating Open Source Geodemographic Classifications for Higher Education Applications. *CASA Working Paper 134: http://www.casa.ucl.ac.uk/working_papers/paper134.pdf.*

Sleight, P. 1997. *Targeting Customers: How to Use Geodeomographic and Lifestyle Data in Your Business.* NTC Publications: Henley on Thames.

Smith, M. T. 2002. Isonymy analysis. The potential for application of quantitative analysis of surname distributions to problems in historical research. In Smith MT (eds.), *Human Biology and History*. 12. Taylor and Francis: London.

Smith, M., Smith, B. and Williams, W. 1984. Changing Isonymic Relationships in Fylingdales Parish, North Yorkshire, 1841-1881. *Annals of Human Biology*. 9: 449-457.

Snae, C. 2007. A Comparison of Name Matching Algorithms. *Proceedings of the World Academy of Science, Engineering and Technology*. 21: 252: 257.

Sokal, R., Harding, R., Lasker. G., Mascie-Taylor, C. 1992. A Spatial Analysis of 100 Surnames in England and Wales. *Annals of Human Biology* 19, 5: 445-476.

Spence, N., Taylor, P. 1970. Quantitative Methods for Regional Taxonomy. *Progress in Geography* 2: 1-64.

Sykes, B., Irven, C. 2000. Surnames and the Y Chromosome. *American Journal of Human Genetics*. 66: 1417- 1419.

Taylor, P, Johnston, R. 1995. GIS and Geography. In Pickles, J. *Ground Truth*. The Guilford Press: New York.

Tobler, W. 1970. A computer movie simulating urban growth in the Detroit region. *Economic geography* 46, 2: 234-241.

Valetas, M, F. 2001. The Surname of Married Women in the European Union. *Bulletin Mensuel D'Inforamation De L'Intstitut National D'Etudes Demographiques*. 367.

Vickers, D., Rees, P. (2007) Creating the UK National Statistics 2001 Output Area Classification. *Journal of the Royal Statistical Society*. Series A. Statistics in Society. 170: 379-403.

Ward, J. 1963. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association* 58, 301:236-244

Webber, R. and Longley, P. 2003. Geodemographic Analysis of Similarity and Proximity: Their Roles in the Understanding of the Geography of Need. In Longley, P. and Batty, M. (eds). *Advanced Spatial Analysis: The CASA Book of GIS*. 233. ESRI Press: Redlands.

Wilson, J. *et al.* 2001. Genetic evidence for different male and female roles during cultural transitions in the British Isles. *Proceedings of the National Academy of Sciences of the United States of America.* 98, 9: 5078-5083.

Winney, B. *et al.* 2011. People of the British Isles: Preliminary Analysis of Genotypes and Surnames in a UK Control Population *European Journal of Human Genetics* (In Press).

Woolland, M., Allen, M. 1999. *1881 Census for England and Wales, the Channel Islands and the Isle of Man: Introductory User Guide V.04*. Distributed by History Data Service, Data Archive, University of Essex, Colchester.

Wrigley, N., Bennett, R. 1981. *Quantitative Geography.* Routledge: London.

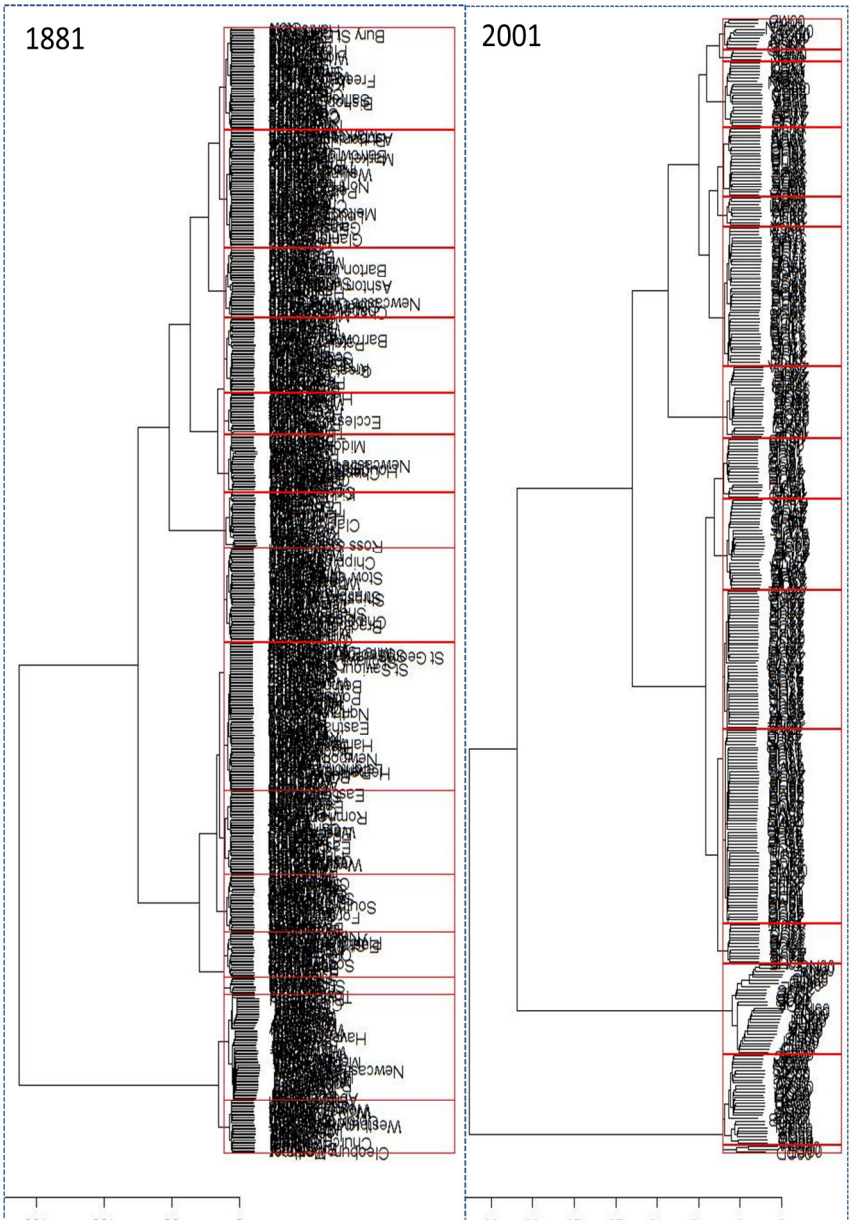
Zelinksy, W. 1970. Cultural Variation in Personal Name Patterns in the Eastern United States. *Annals of the Association of American Geographers.* 60, 4:743-769.

Zelinsky, W. 1997. Along the Frontiers of Name Geography. *Professional Geographer.* 49, 4: 465-466.

10 APPENDIX

1. WARD'S HIERARCHICAL CLUSTERING DENDROGRAMS

Dendrograms illustrating the cophonetic distances between clusters following Ward's clustering of the 1881 (left) and 2001 (right) surnames. The red boxes represent the clusters used to produce Figure 5-13. The first split of the tree distinguishes England and Scotland from Wales in 1881 and England and Wales from Scotland in 2002.



2. CATEGORIES USED TO MAP CELTIC AND VIKING SETTLEMENTS

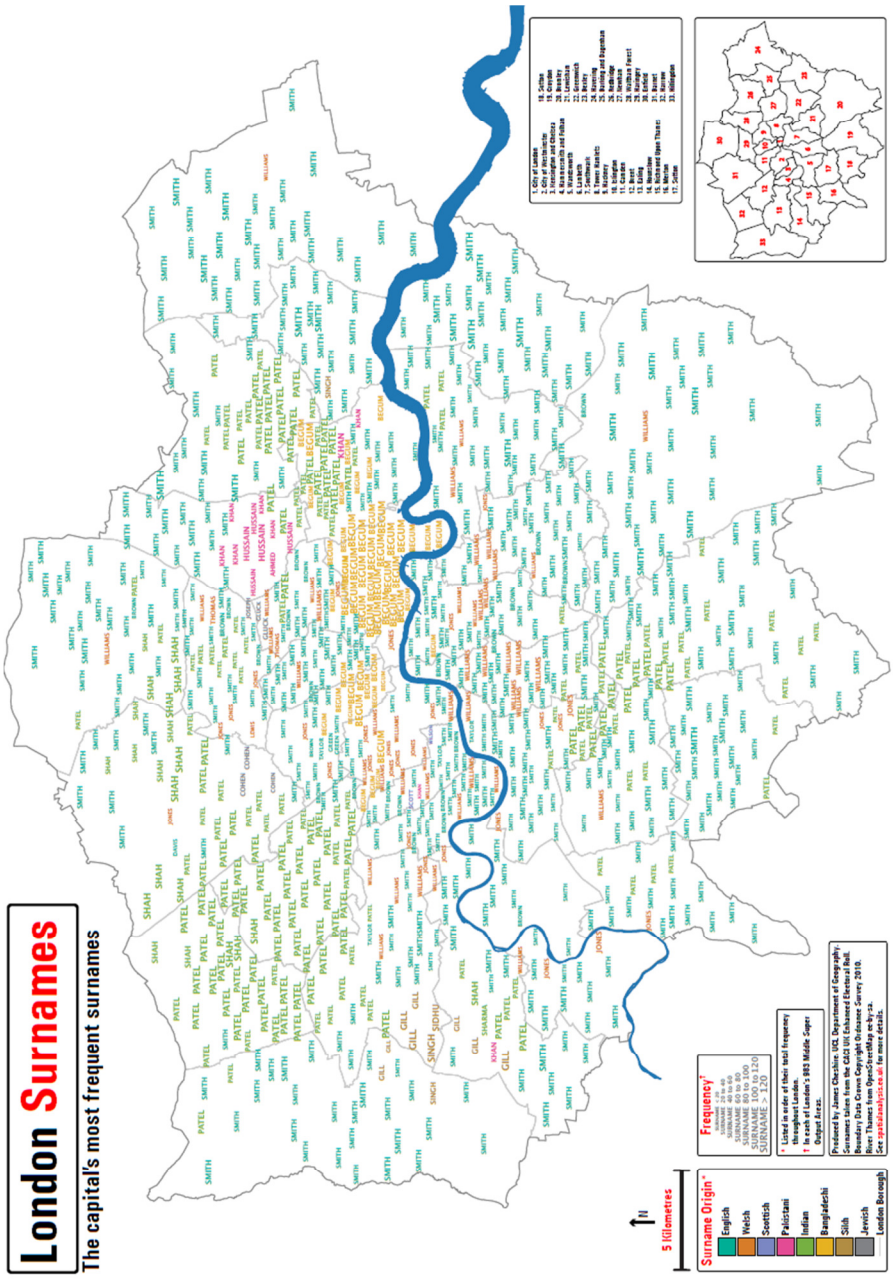
Celtic Naming Conventions:

'aber'
'afon'
'allt'
'don'
'drum'
'brae'
'caer'
'capel'
'coed'
'cwm'
'dinas'
'pont'
'bont'
'porth'
'treath'
'ynys'

Viking/ Danish Conventions:

'thorpe'
'toft'
'holme'
'kirk'
'kir'
'thwaite'
'wick'
'borough'
'ness'

3. LONDON SURNAMES MAP



London Surnames Map (see names.mappinglondon.co.uk for interactive version). Surname origins obtained from the Onomap Classification (see www.onomap.org). The cartography was inspired by a collaborative project mapping the surnames of the USA between National Geographic Magazine and the author. This can be accessed at ngm.nationalgeographic.com/2011/02/geography/usa-surnames-interactive.

4. PUBLISHED JOURNAL PAPERS

Order of inclusion:

2011 Identifying Spatial Concentrations of Surnames. ***International Journal of GIS*** (In Press). (J A Cheshire, P A Longley)*

2011 People of the British Isles: Preliminary Analysis of Genotypes and Surnames in a UK Control Population. ***European Journal of Human Genetics*** (In Press). (B Winney *et al.*)*

2011 Creating a Regional Geography of Britain through the Spatial Analysis of Surnames. ***Geoforum*** (In Press). Doi 10.1016/j.geoforum.2011.02.001 (P A Longley, J A Cheshire, P Mateos)

2010 The Surname Regions of Great Britain. ***Journal of Maps***. Doi 10.4113/jom.2010.1103. (J A Cheshire, P A Longley, A D Singleton)

* Non-proofed versions and therefore subject to minor corrections.